Introduction to IP Multicast Routing

INTERNET-DRAFT C. Semeria T. Maufer
Category: Informational
3Com Corporation January 1997

The following text is Edward Tanguay's edited version of the original document.

FOREWORD

This document is introductory in nature. We have not attempted to describe every detail of each protocol, rather to give a concise overview in all cases, with enough specifics to allow a reader to **grasp the essential details and operation of protocols related to multicast IP**. Every effort has been made to ensure the accurate representation of any cited works, especially any works-in-progress. For the complete details, we refer you to the relevant specification(s).

ABSTRACT

1 The **first** part of this paper describes the **benefits of multicasting, the MBone, Class D addressing, and the operation of the Internet Group Management Protocol (IGMP)**.

2 The second section explores a number of **different techniques** that may potentially be employed by multicast routing protocols:

- Flooding
- Spanning Trees
- Reverse Path Broadcasting (RPB)
- Truncated Reverse Path Broadcasting (TRPB)
- Reverse Path Multicasting (RPM)
- "Shared-Tree" Techniques

3 The third part contains the main body of the paper. It describes **how the previous techniques are implemented in multicast routing protocols available today** (or under development).

- Distance Vector Multicast Routing Protocol (DVMRP)

- Multicast Extensions to OSPF (MOSPF)
- Protocol-Independent Multicast (PIM)
- Core-Based Trees (CBT)

1. Introduction

There are **three** fundamental types of IPv4 addresses: **unicast, broadcast, and multicast.**

1. A **unicast** address is used to transmit a packet to a single destination.
2. A **broadcast** address is used to send a datagram to an entire subnetwork.
3. A **multicast** address is designed to enable the delivery of datagrams to a set of hosts that have been configured as members of a multicast group across various subnetworks.

Multicasting is not connection-oriented. A multicast datagram is delivered to destination group members with the same "best-effort" reliability as a standard unicast IP datagram. **This means that multicast datagrams are not guaranteed to reach all members of a group, nor to arrive in the same order in which they were transmitted.**

The only difference between a multicast IP packet and a unicast IP packet is the presence of a 'group address' in the Destination Address field of the IP header. Instead of a Class A, B, or C IP destination address, **multicasting employs a Class D address format, which ranges from 224.0.0.0 to 239.255.255.255.**

1.1 Multicast Groups

Individual hosts are free to join or leave a multicast group at any time. There are no restrictions on the physical location or the number of members in a multicast group. **A host may be a member of more than one multicast group at any given time** and does not have to belong to a group to send packets to members of a group.

1.2 Group Membership Protocol

A group membership protocol is employed by routers to learn about the presence of group members on their directly attached subnetworks. **When a host joins a multicast group, it transmits a group membership protocol message for the group(s) that it wishes to receive, and sets its IP process and network interface card to receive frames addressed to the multicast group.** This receiver-initiated join process has excellent scaling properties since, as the multicast group increases

in size, it becomes ever more likely that a new group member will be able to locate a nearby branch of the multicast delivery tree.

1.3 Multicast Routing Protocols

Multicast routers execute a multicast routing protocol to define delivery paths that enable the forwarding of multicast datagrams across an internetwork.

1.3.1 Multicast Routing vs. Multicast Forwarding

Multicast routing protocols supply the necessary data to enable the forwarding of multicast packets. In the case of unicast routing, protocols are used to build a forwarding table (commonly called a routing table). Unicast destinations are entered in the routing table, and associated with a metric and a next-hop router toward the destination. Multicast routing protocols are usually unicast routing protocols that **facilitate the determination of routes toward a source, not a destination.** Multicast routing protocols are also used to build a forwarding table.

The key difference between unicast forwarding and multicast forwarding is that multicast packets must be forwarded away from a source. If a packet ever goes back toward the source, a forwarding loop could be formed, possibly leading to a multicast "storm."

A common misconception is that multicast routing protocols pass around information about groups, represented by class D addresses. In fact, as long as a router can determine what direction the source is (relative to itself) and where all the downstream receivers are, then it can build a forwarding table. The forwarding table tells the router that for a certain source sending to a certain group (or in other words, for a certain (source, group) pair), **the packets must all arrive on a certain interface and be copied to certain "downstream" interface(s).**

2. MULTICAST SUPPORT FOR EMERGING INTERNET APPLICATIONS

Today, the **majority of Internet applications rely on point-to-point transmission.** The utilization of point-to-multipoint transmission has traditionally been limited to local area network applications. Over the past few years the Internet has seen a rise in the number of new applications that rely on multicast transmission. Multicast IP conserves bandwidth by forcing the network to do packet replication only when necessary, and offers an attractive alternative to unicast transmission for the delivery of network ticker tapes, live stock quotes, **multiparty videoconferencing**, and **shared whiteboard**

applications (among others). It is important to note that the applications for IP Multicast are not solely limited to the Internet. Multicast IP can also play an important role in large commercial internetworks.

2.1 Reducing Network Load

Assume that a stock ticker application is required to transmit packets to 100 stations within an organization's network. **Unicast transmission to this set of stations will require the periodic transmission of 100 packets where many packets may in fact be traversing the same link(s).** Multicast transmission is the ideal solution for this type of application since it requires only a single packet stream to be transmitted by the source which is replicated at forks in the multicast delivery tree. Unicast retraces routes whereas Multicast streams along the data as it goes.

Broadcast transmission is not an effective solution for this type of application since it affects the CPU performance of each and every station that sees the packet. Besides, it wastes bandwidth.

2.2 Resource Discovery

Some applications implement multicast group addresses instead of broadcasts to transmit packets to group members residing on the same network. However, there is no reason to limit the extent of a multicast transmission to a single LAN. **The time-to-live (TTL) field in the IP header can be used to limit the range (or "scope") of a multicast transmission.**

2.3 Support for Datacasting Applications

@@@

Since 1992, the IETF has conducted a series of "audiocast" experiments in which live audio and video were multicast from the IETF meeting site to destinations around the world. In this case, **"datacasting"** takes compressed audio and video signals from the source station and transmits them as a sequence of UDP packets to a group address. Multicast delivery today is not limited to audio and video. Stock quote systems are one example of a (connectionless) data-oriented multicast application. Someday reliable multicast transport protocols may facilitate efficient inter-computer communication. Reliable multicast transport protocols are currently an active area of research and development.

3. THE INTERNET'S MULTICAST BACKBONE (MBone)

The Internet Multicast Backbone (MBone) is an interconnected set of

subnetworks and routers that support the delivery of IP multicast traffic. The goal of the MBone is to construct a semipermanent IP multicast testbed to enable the deployment of multicast applications without waiting for the ubiquitous deployment of multicast-capable routers in the Internet.

The MBone has grown from 40 subnets in four different countries in 1992, to more than 2800 subnets in over 25 countries by April 1996. With new multicast applications and multicast-based services appearing, it seems

likely that the use of multicast technology in the Internet will keep growing at an ever-increasing rate.

The MBone is a virtual network that is layered on top of sections of the physical Internet. It is composed of islands of multicast routing capability connected to other islands by virtual point-to-point links called "tunnels." The tunnels allow multicast traffic to pass through the non-multicast-capable parts of the Internet. Tunneled IP multicast packets are encapsulated as IP-over-IP (i.e., the protocol number is set to 4) so they look like normal unicast packets to intervening routers. The encapsulation is added on entry to a tunnel and stripped off on exit from a tunnel. This set of multicast routers, their directly-connected subnetworks, and the interconnecting tunnels comprise the MBone.

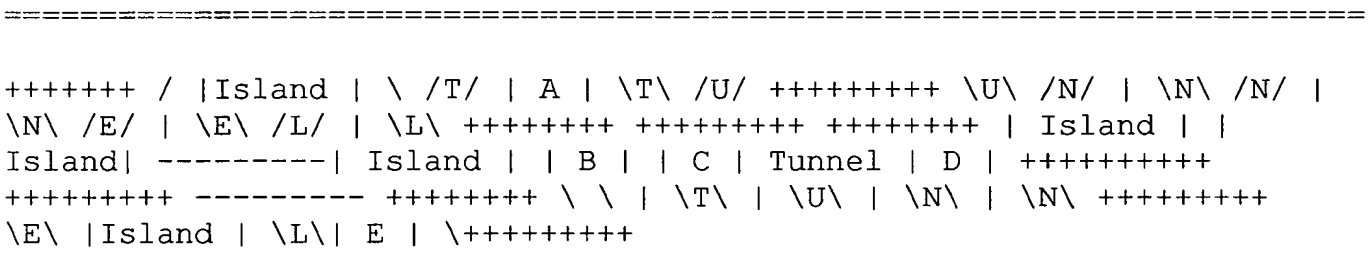


Figure 2: Internet Multicast Backbone (MBone)

Since the MBone and the Internet have different topologies, multicast routers execute a separate routing protocol to decide how to forward multicast packets. The majority of the MBone routers currently use the Distance Vector Multicast Routing Protocol (DVMRP), although some portions of the MBone execute either Multicast OSPF (MOSPF) or the Protocol-Independent Multicast (PIM) routing protocols. The operation of each of these protocols is discussed later in this paper.

As multicast routing software features become more widely available on the routers of the Internet, providers may gradually decide to use "native" multicast as an alternative to using lots of tunnels.

The MBone carries audio and video multicasts of Internet Engineering Task Force (IETF) meetings, NASA Space Shuttle Missions, US House and Senate sessions, and live satellite weather photos. The session directory (SDR) tool provides users with a listing of the active multicast sessions on the MBone and allows them to create and/or join a session.

4. MULTICAST ADDRESSING

A multicast address is assigned to a set of receivers defining a multicast group. Senders use the multicast address as the destination IP address of a packet that is to be transmitted to all group members.

4.1 Class D Addresses

An IP multicast group is identified by a Class D address. Class D addresses have their high-order four bits set to "1110" followed by a 28-bit multicast group ID. Expressed in standard "dotted-decimal" notation, multicast group addresses range from 224.0.0.0 to 239.255.255.255 (shorthand: 224.0.0.0/4).

Figure 3 shows the format of a 32-bit Class D address.

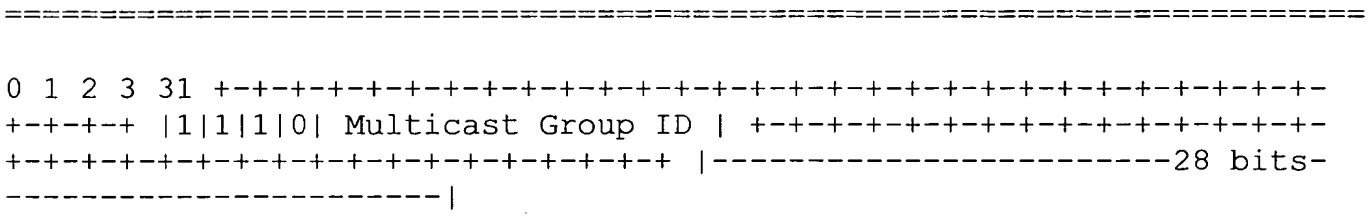


Figure 3: Class D Multicast Address Format

The Internet Assigned Numbers Authority (IANA) maintains a list of registered IP multicast groups. The base address 224.0.0.0 is reserved and cannot be assigned to any group. The block of multicast addresses ranging from 224.0.0.1 to 224.0.0.255 is reserved for permanent assignment to various uses, including routing protocols and other protocols that require a well-known permanent address. Multicast routers should not forward any multicast datagram with

destination addresses in this range, (regardless of the packet's TTL).

Some of the well-known groups include:

"all systems on this subnet" 224.0.0.1 "all routers on this subnet" 224.0.0.2 "all DVMRP routers" 224.0.0.4 "all OSPF routers" 224.0.0.5 "all OSPF designated routers" 224.0.0.6 "all RIP2 routers" 224.0.0.9 "all PIM routers" 224.0.0.13

The remaining groups ranging from 224.0.1.0 to 239.255.255.255 are assigned to various multicast applications or remain unassigned. From this range, the addresses from 239.0.0.0 to 239.255.255.255 are being reserved for site-local "administratively scoped" applications, not Internet-wide applications.

The complete list may be found in the Assigned Numbers RFC (RFC 1700 or its successor) or at the IANA Web Site:

<URL:http://www.isi.edu/div7/iana/assignments.html>

4.2 Mapping a Class D Address to an IEEE-802 MAC Address

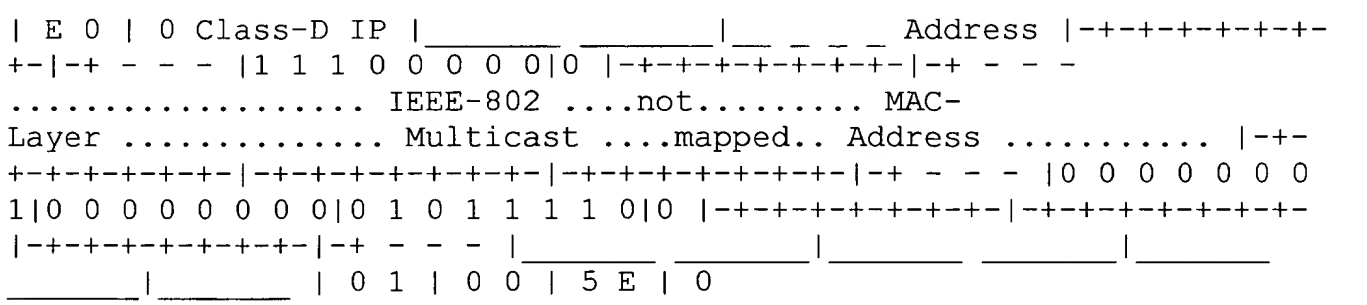
The IANA has been allocated a reserved portion of the IEEE-802 MAC-layer multicast address space. All of the addresses in IANA's reserved block begin with 01-00-5E (hex). A simple procedure was developed to map Class D addresses to this reserved address block. This allows IP multicasting to easily take advantage of the hardware-level multicasting supported by network interface cards.

For example, the mapping between a Class D IP address and an IEEE-802 (e.g., Ethernet) multicast address is obtained by placing the low-order 23 bits of the Class D address into the low-order 23 bits of IANA's reserved address block.

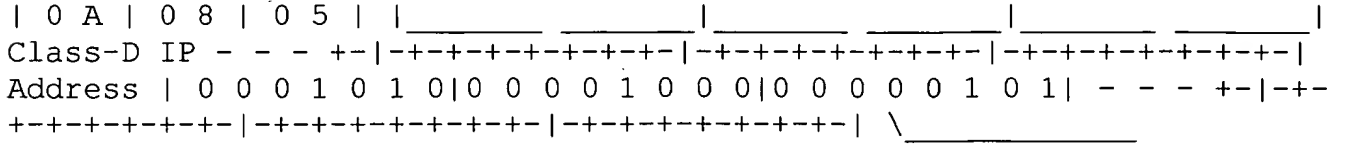
Figure 4 illustrates how the multicast group address 224.10.8.5 (E0-0A-08-05) is mapped into an IEEE-802 multicast address.

==========

Class D Address: 224.10.8.5 (E0-0A-08-05)



[Address mapping below continued from half above]



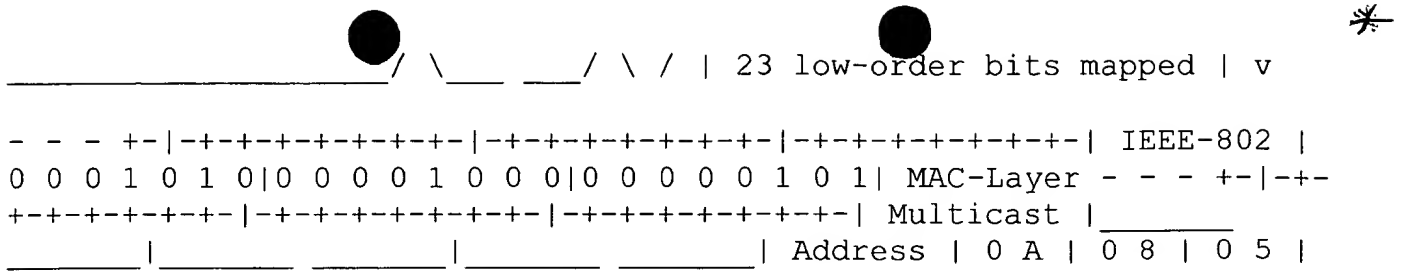


Figure 4: Mapping between Class D and IEEE-802 Multicast Addresses

The mapping in Figure 4 places the low-order 23 bits of the IP multicast group ID into the low order 23 bits of the IEEE-802 multicast address. Note that the mapping may place up to 32 different IP groups into the

same IEEE-802 address because the upper 5 bits of the IP multicast group ID are not used. For example, the multicast addresses 224.138.8.5 (E0-8A-08-05) and 225.10.8.5 (E1-0A-08-05) would also be mapped to the same IEEE-802 multicast address (01-00-5E-0A-08-05) used in this example.

4.3 Transmission and Delivery of Multicast Datagrams

When the sender and receivers are members of the same (LAN) subnetwork, the transmission and reception of multicast frames is a straightforward process. The source station simply addresses the IP packet to the multicast group, the network interface card maps the Class D address to the corresponding IEEE-802 multicast address, and the frame is sent. Receivers that wish to capture the frame notify their MAC and IP layers that they want to receive datagrams addressed to the group.

Things become somewhat more complex when the sender is attached to one subnetwork and receivers reside on different subnetworks. In this case, the routers must implement a multicast routing protocol that permits the construction of multicast delivery trees and supports multicast packet forwarding. In addition, each router needs to implement a group membership protocol that allows it to learn about the existence of group members on its directly attached subnetworks.

5. INTERNET GROUP MANAGEMENT PROTOCOL (IGMP)

The Internet Group Management Protocol (IGMP) runs between hosts and their immediately-neighboring multicast routers. The mechanisms of the protocol allow a host to inform its local router that it wishes

to receive transmissions addressed to a specific multicast group. Also, routers periodically query the LAN to determine if known group members are still active. If there is more than one router on the LAN performing IP multicasting, one of the routers is elected "querier" and assumes the responsibility of querying the LAN for group members.

Based on the group membership information learned from the IGMP, a router is able to determine which (if any) multicast traffic needs to be forwarded to each of its "leaf" subnetworks. Multicast routers use this information, in conjunction with a multicast routing protocol, to support IP multicasting across the Internet.

5.1 IGMP Version 1

IGMP Version 1 was specified in RFC-1112. According to the specification, multicast routers periodically transmit Host Membership Query messages to determine which host groups have members on their directlyattached networks. Query messages are addressed to the all-hosts group (224.0.0.1) and have an IP TTL = 1. This means that Query messages sourced from a router are transmitted onto the

directly-attached subnetwork but are not forwarded by any other multicast routers.

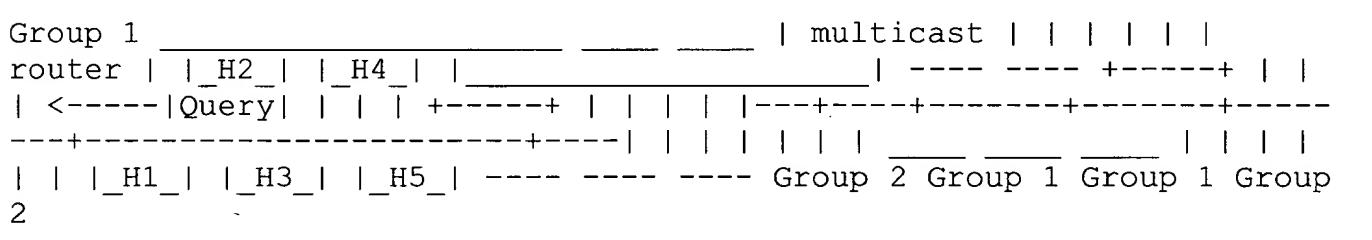


Figure 5: Internet Group Management Protocol-Query Message

When a host receives an IGMP Query message, it responds with a Host

Membership Report for each group to which it belongs, sent to each group to which it belongs. (This is an important point: While IGMP Queries are sent to the "all hosts on this subnet" class D address (224.0.0.1), IGMP Reports are sent to the group(s) to which the host (s) belong. Reports have a TTL of 1, and thus are not forwarded beyond the local subnetwork.)

In order to avoid a flurry of Reports, each host starts a randomly-chosen Report delay timer for each of its group memberships. If, during the delay period, another Report is heard for the same group, each other host in that group resets its timer to a new random value. This procedure spreads Reports out over a period of time and minimizes Report traffic for each group that has at least one member on a given subnetwork.

It should be noted that multicast routers do not need to be directly addressed since their interfaces are required to promiscuously receive all multicast IP traffic. Also, a router does not need to maintain a detailed list of which hosts belong to each multicast group; the router

only needs to know that at least one group member is present on a given network interface.

Multicast routers periodically transmit Queries to update their knowledge of the group members present on each network interface. If the router does not receive a Report from any members of a particular group after a number of Queries, the router assumes that group members are no longer present on an interface. Assuming this is a leaf subnet, this interface is removed from the delivery tree for this (source, group) pair. Multicasts will continue to be sent on this interface if the router can tell (via multicast routing protocols) that there are additional group members further downstream reachable via this interface.

When a host first joins a group, it immediately transmits an IGMP Report for the group rather than waiting for a router's IGMP Query. This reduces the "join latency" for the first host to join a given group on a particular subnetwork.

5.2 IGMP Version 2

IGMP Version 2 was distributed as part of the IP Multicasting (Version 3.3 through Version 3.8) code package. Initially, there was no detailed specification for IGMP Version 2 other than this source

code. However, the complete specification has recently been published in <draft-ietf- idmr-igmp-v2-05.txt> which will update the informal specification contained in Appendix I of RFC-1112. IGMP Version 2 enhances and extends IGMP Version 1 while maintaining backward compatibility with Version 1 hosts.

IGMP Version 2 defines a procedure for the election of the multicast querier for each LAN. In IGMP Version 2, the router with the lowest IP address on the LAN is elected the multicast querier. In IGMP Version 1, the querier election was determined by the multicast routing protocol. This could lead to potential problems because each multicast routing protocol might use unique methods for determining the multicast querier.

IGMP Version 2 defines a new type of Query message: the Group-Specific Query. Group-Specific Query messages allow a router to transmit a Query to a specific multicast group rather than all groups residing on a directly attached subnetwork.

Finally, IGMP Version 2 defines a Leave Group message to lower IGMP's "leave latency." When the last host to respond to a Query with a Report wishes to leave that specific group, the host transmits a Leave Group message to the all-routers group (224.0.0.2) with the group field set to the group to be left. In response to a Leave Group message, the router begins the transmission of Group-Specific Query messages on the inter- face that received the Leave Group message. If there are no Reports in response to the Group-Specific Query messages, then if this is a leaf

subnet, this interface is removed from the delivery tree for this (source, group) pair (as was the case of IGMP version 1). Again, multicasts will continue to be sent on this interface if the router can tell (via multicast routing protocols) that there are additional group members further downstream reachable via this interface.

5.3 IGMP Version 3

IGMP Version 3 is a preliminary draft specification published in <draft-cain-igmp-00.txt>. IGMP Version 3 introduces support for Group- Source Report messages so that a host can elect to receive traffic from specific sources of a multicast group. An Inclusion

Group-Source Report message allows a host to specify the IP addresses of the specific sources it wants to receive. An Exclusion Group-Source Report message allows a host to explicitly identify the sources that it does not want to receive. With IGMP Version 1 and Version 2, if a host wants to receive any traffic for a group, the traffic from all sources for the group must be forwarded onto the host's subnetwork.

IGMP Version 3 will help conserve bandwidth by allowing a host to select the specific sources from which it wants to receive traffic. Also, multicast routing protocols will be able to make use this information to conserve bandwidth when constructing the branches of their multicast delivery trees.

Finally, support for Leave Group messages first introduced in IGMP Version 2 has been enhanced to support Group-Source Leave messages. This feature allows a host to leave an entire group or to specify the specific IP address(es) of the (source, group) pair(s) that it wishes to leave.

6. MULTICAST FORWARDING TECHNIQUES

IGMP provides the final step in a multicast packet delivery service since it is only concerned with the forwarding of multicast traffic from a router to group members on its directly-attached subnetworks. IGMP is not concerned with the delivery of multicast packets between neighboring routers or across an internetwork.

To provide an internetwork delivery service, it is necessary to define multicast routing protocols. A multicast routing protocol is responsible for the construction of multicast delivery trees and enabling multicast packet forwarding. This section explores a number of

different techniques that may potentially be employed by multicast routing protocols:

o "Simpleminded" Techniques - Flooding - Spanning Trees

o Source-Based Tree (SBT) Techniques - Reverse Path Broadcasting (RPB) - Truncated Reverse Path Broadcasting (TRPB) - Reverse Path Multicasting (RPM) o "Shared-Tree" Techniques

Later sections will describe how these algorithms are implemented in the most prevalent multicast routing protocols in the Internet today (e.g., Distance Vector Multicast Routing Protocol (DVMRP), Multicast extensions to OSPF (MOSPF), Protocol-Independent Multicast (PIM), and Core-Based Trees (CBT).

6.1 "Simpleminded" Techniques

Flooding and Spanning Trees are two algorithms that can be used to build primitive multicast routing protocols. The techniques are primitive due to the fact that they tend to waste bandwidth or require a large amount of computational resources within the multicast routers involved. Also, protocols built on these techniques may work for small networks with few senders, groups, and routers, but do not scale well to larger numbers of senders, groups, or routers. Also, the ability to handle arbitrary topologies may not be present or may only be present in limited ways.

6.1.1 Flooding

The simplest technique for delivering multicast datagrams to all routers in an internetwork is to implement a flooding algorithm. The flooding procedure begins when a router receives a packet that is addressed to a multicast group. The router employs a protocol mechanism to determine whether or not it has seen this particular packet before. If it is the first reception of the packet, the packet is forwarded on all interfaces--except the one on which it arrived--guaranteeing that the multicast packet reaches all routers in the internetwork. If the router has seen the packet before, then the packet is discarded.

A flooding algorithm is very simple to implement since a router does not have to maintain a routing table and only needs to keep track of the most recently seen packets. However, flooding does not scale for Internet-wide applications since it generates a large number of duplicate packets and uses all available paths across the

internetwork instead of just a limited number. Also, the flooding algorithm makes inefficient use of router memory resources since each router is required to maintain a distinct table entry for each recently seen packet.

6.1.2 Spanning Tree

A more effective solution than flooding would be to select a subset of the internetwork topology which forms a spanning tree. The spanning tree defines a structure in which only one active path connects any two routers of the internetwork. Figure 6 shows an internetwork and a spanning tree rooted at router RR.

Once the spanning tree has been built, a multicast router simply forwards each multicast packet to all interfaces that are part of the spanning tree except the one on which the packet originally arrived. Forwarding along the branches of a spanning tree guarantees that the multicast packet will not loop and that it will eventually reach all routers in the internetwork.

A spanning tree solution is powerful and would be relatively easy to implement since there is a great deal of experience with spanning tree protocols in the Internet community. However, a spanning tree solution can centralize traffic on a small number of links, and may not provide the most efficient path between the source subnetwork and group members. Also, it is computationally difficult to compute a spanning tree in large, complex topologies.

6.2 Source-Based Tree Techniques

The following techniques all generate a source-based tree by various means. The techniques differ in the efficiency of the tree building process, and the bandwidth and router resources (i.e., state tables) used to build a source-based tree.

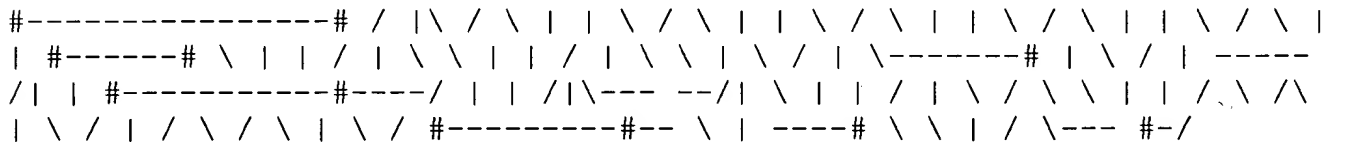
6.2.1 Reverse Path Broadcasting (RPB)

A more efficient solution than building a single spanning tree for the entire internetwork would be to build a group-specific spanning tree for each potential source [subnetwork]. These spanning trees would result in source-based delivery trees emanating from the subnetwork directly connected to the source station. Since there are many potential sources for a group, a different delivery tree is constructed emanating from each active source.

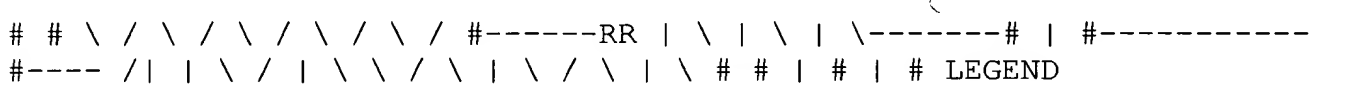
6.2.1.1 Reverse Path Broadcasting: Operation

The fundamental algorithm to construct these source-based trees is referred to as Reverse Path Broadcasting (RPB). The RPB algorithm is actually quite simple. For each (source, group) pair, if a packet arrives on a link that the local router believes to be on the shortest path back toward the packet's source, then the router forwards the packet on all interfaces except the incoming interface. If the packet does not arrive on the interface that is on the shortest path back

A Sample Internetwork



A Spanning Tree for this Sample Internetwork



Router RR Root Router

Figure 6: Spanning Tree

toward the source, then the packet is discarded. The interface over which the router expects to receive multicast packets from a particular source is referred to as the "parent" link. The outbound links over which the router forwards the multicast packet are called "child" links for this group.

This basic algorithm can be enhanced to reduce unnecessary packet

duplication. If the local router making the forwarding decision can determine whether a neighboring router on a child link is "downstream," then the packet is multicast toward the neighbor. (A "downstream" neighbor is a neighboring router which considers the local router to be on the shortest path back toward a given source.) Otherwise, the packet is not forwarded on the potential child link since the local router knows that the neighboring router will just discard the packet (since it will arrive on a non-parent link for the (source, group) pair, relative to that downstream router).

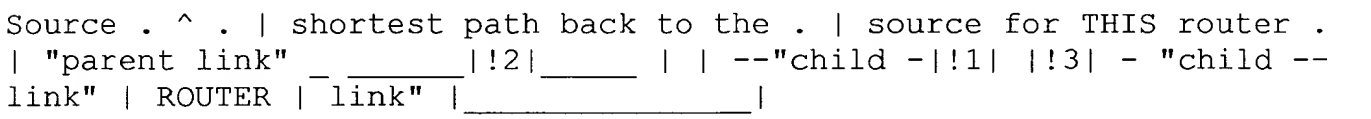
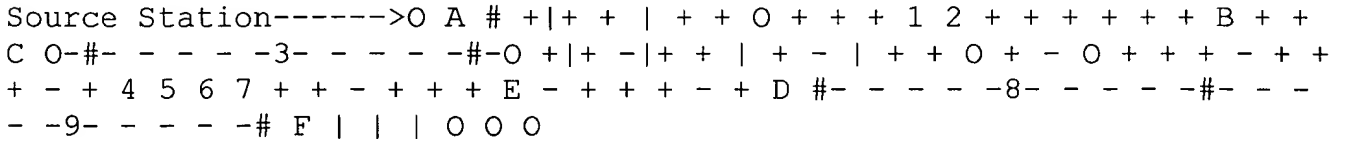


Figure 7: Reverse Path Broadcasting - Forwarding Algorithm

The information to make this "downstream" decision is relatively easy to derive from a link-state routing protocol since each router maintains a topological database for the entire routing domain. If a distance-vector routing protocol is employed, a neighbor can either advertise its previous hop for the (source, group) pair as part of its routing update messages or "poison reverse" the route toward a source if it is not on the distribution tree for that source. Either of these techniques allows an upstream router to determine if a downstream neighboring router is on an active branch of the delivery tree for a certain source sending to a certain group.

Please refer to Figure 8 for a discussion describing the basic operation of the enhanced RPB algorithm.



LEGEND

O Leaf + + Shortest path - - Branch # Router

Figure 8: Reverse Path Broadcasting - Example

Note that the source station (S) is attached to a leaf subnetwork directly connected to Router A. For this example, we will look at the RPB algorithm from Router B's perspective. Router B receives the multicast packet from Router A on link 1. Since Router B considers link 1 to be the parent link for the (source, group) pair, it forwards the packet on link 4, link 5, and the local leaf subnetworks if they contain group members. Router B does not forward the packet on link 3 because

it knows from routing protocol exchanges that Router C considers link 2 as its parent link for the (source, group) pair. Router B knows that if it were to forward the packet on link 3, it would be discarded by Router C since the packet would not be arriving on Router C's parent link for this (source, group) pair.

6.2.1.2 RPB: Benefits and Limitations

The key benefit to reverse path broadcasting is that it is reasonably efficient and easy to implement. It does not require that the router know about the entire spanning tree, nor does it require a special mechanism to stop the forwarding process (as flooding does). In addition, it guarantees efficient delivery since multicast packets always follow the "shortest" path from the source station to the destination group. Finally, the packets are distributed over multiple links, resulting in better network utilization since a different tree is computed for each (source, group) pair.

One of the major limitations of the RPB algorithm is that it does not take into account multicast group membership when building the delivery tree for a (source, group) pair. As a result, datagrams may be unnecessarily forwarded to subnetworks that have no members in the destination group.

6.2.2 Truncated Reverse Path Broadcasting (TRPB)

Truncated Reverse Path Broadcasting (TRPB) was developed to overcome the limitations of Reverse Path Broadcasting. With the help of IGMP, multicast routers determine the group memberships on each leaf subnetwork and avoid forwarding datagrams onto a leaf subnetwork if it does not contain at least one member of the destination group.

Thus, the delivery tree is "truncated" by the router if a leaf subnetwork has no group members.

Figure 9 illustrates the operation of TRPB algorithm. In this example the router receives a multicast packet on its parent link for the (Source, G1) pair. The router forwards the datagram on interface 1 since that interface has at least one member of G1. The router does not forward the datagram to interface 3 since this interface has no members in the destination group. The datagram is forwarded on interface 4 if and only if a downstream router considers this subnetwork to be part of its "parent link" for the (Source, G1) pair.

TRPB removes some limitations of RPB but it solves only part of the problem. It eliminates unnecessary traffic on leaf subnetworks but it does not consider group memberships when building the branches of the delivery tree.

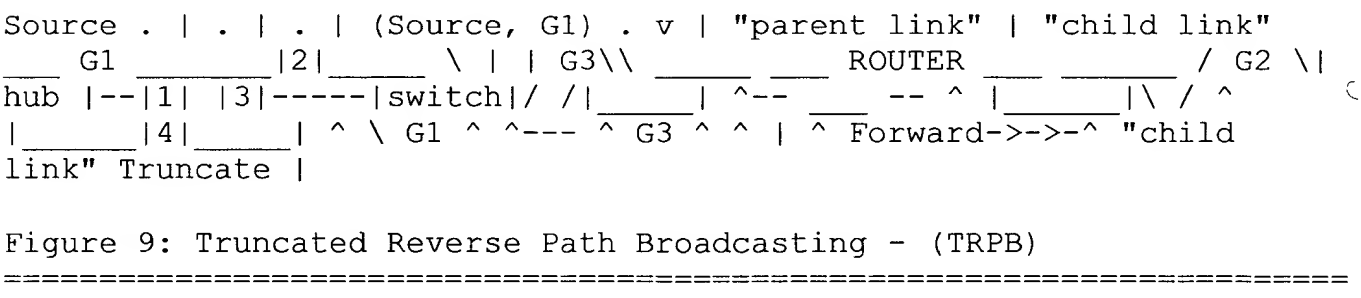


Figure 9: Truncated Reverse Path Broadcasting - (TRPB)

6.2.3 Reverse Path Multicasting (RPM)

Reverse Path Multicasting (RPM) is an enhancement to Reverse Path Broadcasting and Truncated Reverse Path Broadcasting.

RPM creates a delivery tree that spans only:

o Subnetworks with group members, and

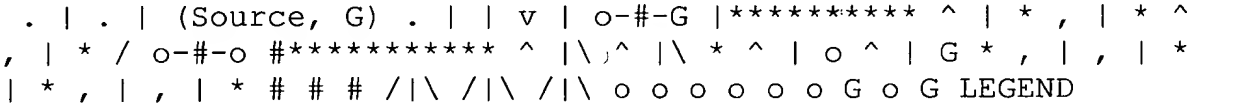
o Routers and subnetworks along the shortest path to subnetworks with group members.

RPM allows the source-based "shortest-path" tree to be pruned so that datagrams are only forwarded along branches that lead to active members of the destination group.

6.2.3.1 Operation

When a multicast router receives a packet for a (source, group) pair, the first packet is forwarded following the TRPB algorithm across all routers in the internetwork. Routers on the edge of the network (which have only leaf subnetworks) are called leaf routers. The TRPB algorithm guarantees that each leaf router will receive at least the first multicast packet. If there is a group member on one of its leaf

subnetworks, a leaf router forwards the packet based on this IGMP Report (or a statically-defined local group on an interface).



Router o Leaf without group member G Leaf with group member *** Active Branch --- Pruned Branch ,>, Prune Message (direction of flow -->

Figure 10: Reverse Path Multicasting (RPM)

If none of the subnetworks connected to the leaf router contain group members, the leaf router may transmit a "prune" message on its parent link, informing the upstream router that it should not forward packets for this particular (source, group) pair on the child interface on which it received the prune message. Prune messages are sent just one hop back toward the source.

An upstream router receiving a prune message is required to store the prune information in memory. If the upstream router has no recipients

on local leaf subnetworks and has received prune messages on each of the child interfaces for this (source, group) pair, then the upstream router

does not need to receive additional packets for the (source, group) pair. This implies that the upstream router can also generate a prune message of its own, one hop further back toward the source. This cascade of prune messages results in an active multicast delivery tree, consisting exclusively of "live" branches (i.e., branches that lead to active receivers).

Since both the group membership and internetwork topology can change dynamically , the pruned state of the multicast delivery tree must be refreshed periodically. At regular intervals, the prune information expires from the memory of all routers and the next packet for the (source, group) pair is forwarded toward all downstream routers. This results in a new burst of prune messages allowing the multicast forwarding tree to adapt to the ever-changing multicast delivery requirements of the internetwork.

6.2.3.2 Limitations

Despite the improvements offered by the RPM algorithm, there are still several scaling issues that need to be addressed when attempting to develop an Internet-wide delivery service. The first limitation is that multicast packets must be periodically flooded across every router in the internetwork, onto every leaf subnetwork. This flooding is wasteful of bandwidth (until the updated prune state is constructed).

This "flood and prune" paradigm is very powerful, but it wastes bandwidth and does not scale well, especially if there are receivers at the edge of the delivery tree which are connected via low-speed technologies (e.g., ISDN or modem). Also, note that every router participating in the RPM algorithm must either have a forwarding table entry for a (source, group) pair, or have prune state information for that (source, group) pair.

It is clearly wasteful (especially as the number of active sources and groups increase) to place such a burden on routers that are not on every (or perhaps any) active delivery tree. Shared tree techniques are an attempt to address these scaling issues, which become quite acute when most groups' senders and receivers are sparsely distributed across the internetwork.

6.3 Shared Tree Techniques

The most recent additions to the set of multicast forwarding techniques are based on a shared delivery tree. Unlike shortest-path tree algorithms which build a source-based tree for each (source, group) pair, shared tree algorithms construct a single delivery tree that is shared by all members of a group. The shared tree approach is quite similar to the spanning tree algorithm except it allows the definition of a different shared tree for each group. Stations that wish to receive traffic for a multicast group are required to explicitly join

the shared delivery tree. Multicast traffic for each group is sent and received over the same delivery tree, regardless of the source.

6.3.1 Operation

A shared tree may involve a single router, or set of routers, which comprise the "core" of a multicast delivery tree. Figure 11 illustrates how a single multicast delivery tree is shared by all sources and receivers for a multicast group.

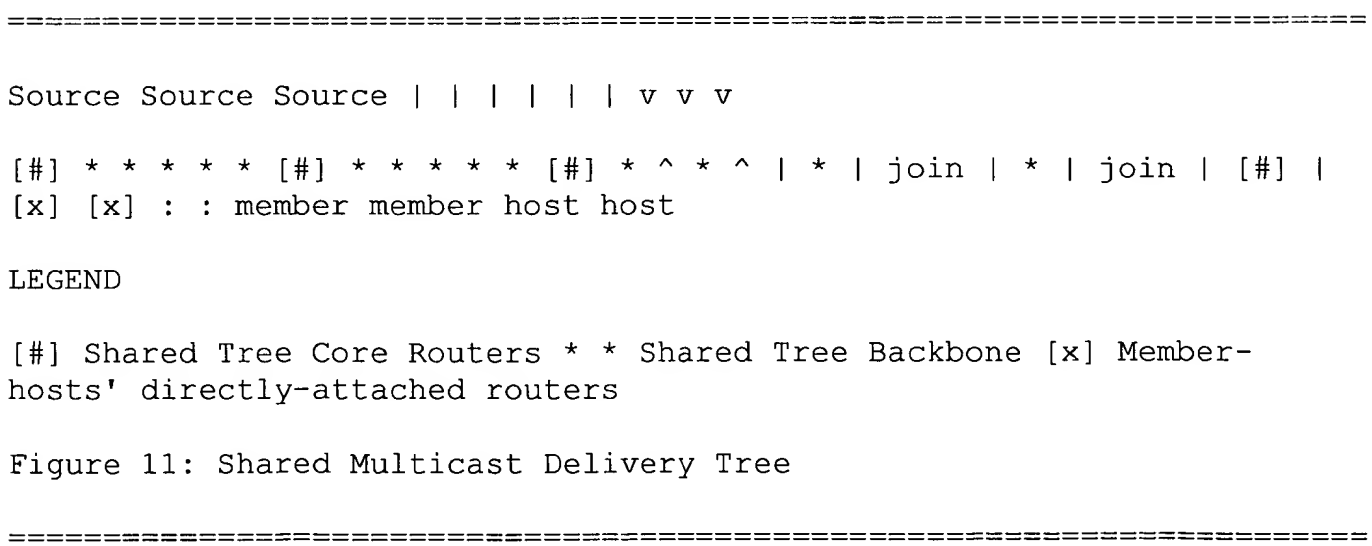


Figure 11: Shared Multicast Delivery Tree

The directly attached router for each station wishing to belong to a particular multicast group is required to send a "join" message toward the shared tree of the particular multicast group. The

directly attached router only needs to know the address of one of the group's core routers in order to transmit a join request (via unicast). The join request is processed by all intermediate routers, each of which identifies the interface on which the join was received as belonging to the group's delivery tree. The intermediate routers continue to forward the join message toward the core, marking local downstream interfaces

until the request reaches a core router (or a router that is already on the active delivery tree). This procedure allows each member-host's directly-attached router to define a branch providing the shortest path between itself and a core router which is part of the group's shared delivery tree.

Similar to other multicast forwarding algorithms, shared tree algorithms do not require that the source of a multicast packet be a member of a destination group. Packets sourced by a non-group member are simply unicast toward the core until they reach the first router that is a member of the group's delivery tree. When the unicast packet reaches a member of the delivery tree, the packet is multicast to all outgoing interfaces that are part of the tree except the incoming link. This guarantees that traffic follows the shortest path from source station to the shared tree. It also ensures that multicast packets are forwarded to all routers on the core tree which in turn forward the traffic to all receivers that have joined the shared tree.

6.3.2 Benefits

In terms of scalability, shared tree techniques have several advantages over source-based trees. Shared tree algorithms make efficient use of router resources since they only require a router to maintain state information for each group, not for each (source, group) pair. (Remember that source-based tree techniques required all routers in an internetwork to either be a) on the delivery tree for a given (source, group) pair, or b) have prune state for that (source, group) pair: So the entire internetwork must participate in the source-based tree protocol.) This improves the scalability of applications with many active senders since the number of source stations is no longer a scaling issue. Also, shared tree algorithms conserve network bandwidth since they do not require that multicast packets be periodically flooded across all multicast routers in the internetwork onto every leaf subnetwork. This can offer significant bandwidth savings, especially across low-bandwidth WAN links, and when receivers sparsely populate the domain of operation. Finally, since receivers are required to explicitly join the shared delivery

tree, data only ever flows over those links that lead to active receivers.

6.3.3 Limitations

Despite these benefits, there are still several limitations to protocols that are based on a shared tree algorithm. Shared trees may result in traffic concentration and bottlenecks near core routers since traffic from all sources traverses the same set of links as it approaches the core. In addition, a single shared delivery tree may create suboptimal routes (a shortest path between the source and the shared tree, a suboptimal path across the shared tree, a shortest path between the egress core router and the receiver's directly attached router) resulting in increased delay which may be a critical issue for some multimedia applications. (Simulations indicate that latency over a

shared tree may be approximately 10% larger than source-based trees in many cases, but by the same token, this may be negligible for many applications.) Finally, new algorithms need to be developed to support all aspects of core management which include core router selection and (potentially) dynamic placement strategies.

7. SOURCE-BASED TREE ("DENSE MODE") ROUTING PROTOCOLS

An established set of multicast routing protocols define a source-based delivery tree which provides the shortest path between the source and each receiver.

These routing protocols include:

o Distance Vector Multicast Routing Protocol (DVMRP),

o Multicast Extensions to Open Shortest Path First (MOSPF),

o Protocol Independent Multicast - Dense Mode (PIM-DM).

Each of these routing protocols is designed to operate in an environment where group members are relatively densely populated and internetwork bandwidth is plentiful. Their underlying designs assume that the amount of protocol overhead (in terms of the amount of state that must be maintained by each router, the number of router CPU cycles required, and the amount of bandwidth consumed by protocol operation) is appropriate since receivers densely populate the area of operation.

7.1. Distance Vector Multicast Routing Protocol (DVMRP)

The Distance Vector Multicast Routing Protocol (DVMRP) is a distance-vector routing protocol designed to support the forwarding of multicast datagrams through an internetwork. DVMRP constructs source-based multicast delivery trees using variants of the Reverse Path Broadcasting (RPB) algorithm. Originally, the entire MBone ran DVMRP. Today, over half of the MBone routers still run some version of DVMRP.

DVMRP was first defined in RFC-1075. The original specification was derived from the Routing Information Protocol (RIP) and employed the Truncated Reverse Path Broadcasting (TRPB) technique. The major difference between RIP and DVMRP is that RIP was concerned with calculating the next-hop to a destination, while DVMRP is concerned with computing the previous-hop back to a source. It is important to note that the latest mrouted version 3.8 and vendor implementations have extended DVMRP to employ the Reverse Path Multicasting (RPM) algorithm. This means that the latest implementations of DVMRP are quite different from the original RFC specification in many regards. There is an active effort within the IETF Inter-Domain Multicast Routing (IDMR) working group to specify DVMRP version 3 in a standard form (as opposed to the current spec, which is written in C).

The current DVMRP v3 Internet-Draft is:

<draft-ietf-idmr-dvmrp-v3-03.txt>, or <draft-ietf-idmr-dvmrp-v3-03.ps>

7.1.1 Physical and Tunnel Interfaces

The ports of a DVMRP router may be either a physical interface to a directly-attached subnetwork or a tunnel interface to another multicast island. All interfaces are configured with a metric that specifies the cost for the given port and a TTL threshold that limits the scope of a multicast transmission. In addition, each tunnel interface must be explicitly configured with two additional parameters: the IP address of the local router's interface and the IP address of the remote router's interface.

TTL Scope Threshold
0 Restricted to the same host 1 Restricted to the same subnetwork 15

Restricted to the same site 63 Restricted to the same region 127
Worldwide 191 Worldwide; limited bandwidth 255 Unrestricted in scope

Table 1: TTL Scope Control Values

A multicast router will only forward a multicast datagram across an interface if the TTL field in the IP header is greater than the TTL threshold assigned to the interface. Table 1 lists the conven- tional TTL values that are used to restrict the scope of an IP multicast. For example, a multicast datagram with a TTL of less than 16 is restricted to the same site and should not be forwarded across an interface to other sites in the same region.

TTL-based scoping is not always useful, so the IETF MBoneD working group is considering the definition and usage of a range of multi- cast addresses for "administrative" scoping. In other words, such addresses would be usable within a certain administrative scope, a corporate network, for instance, but would not be forwarded across the global MBone. At the moment, the range from 239.0.0.0 through 239.255.255.255 is being reserved for administratively scoped applications, but the structure and usage of this block has yet to be completely formalized.

7.1.2 Basic Operation

DVMRP implements the Reverse Path Multicasting (RPM) algorithm. According to RPM, the first datagram for any (source, group) pair is forwarded across the entire internetwork (providing the packet's TTL and router interface thresholds permit this). The initial datagram is delivered to all leaf routers which transmit prune messages back toward the source if there are no group members on their directly attached leaf subnetworks. The prune messages result in the removal of branches from the tree that do not lead to group members, thus creating a source-based shortest path tree with all leaves having group members. After a period of time, the pruned branches grow back and the next datagram for the (source, group) pair is forwarded across the entire internetwork resulting in a new set of prune messages.

DVMRP also implements a mechanism to quickly "graft" back a previously pruned branch of a group's delivery tree. If a router that previously sent a prune message for a (source, group) pair discovers new group members on a leaf network, it sends a graft message to the group's previous-hop router. When an upstream router receives a graft

message, it cancels out the previously-received prune message. Graft messages will cascade back toward the source (until reaching the nearest "live" branch point on the delivery tree), thus allowing previously pruned branches to be quickly restored as part of the active delivery tree.

7.1.3 DVMRP Router Functions

When there is more than one DVMRP router on a subnetwork, the Dominant Router is responsible for the periodic transmission of IGMP Host Membership Query messages. Upon initialization, a DVMRP router considers itself to be the Dominant Router for the subnetwork until it receives a Host Membership Query message from a neighbor router with a lower IP address. Figure 12 illustrates how the router with the lowest IP address functions as the Dominant Router for the subnetwork.

In order to avoid duplicate multicast datagrams when there is more than one DVMRP router on a subnetwork, one router is elected the Dominant Router for the particular source subnetwork (see fig. 12). In Figure 13, Router C is downstream and may potentially receive datagrams from the source subnetwork from Router A or Router B. If Router A's metric to the source subnetwork is less than Router B's metric, then Router A is dominant over Router B for this source.

This means that Router A will forward traffic from the source subnetwork and Router B will discard traffic from that source subnetwork. However, if Router A's metric is equal to Router B's metric, then router with the lower IP address on its downstream interface (child link) becomes the Dominant Router for this source. Note that on a subnetwork with multiple routers forwarding to groups with multiple sources, different routers may be dominant for each source.

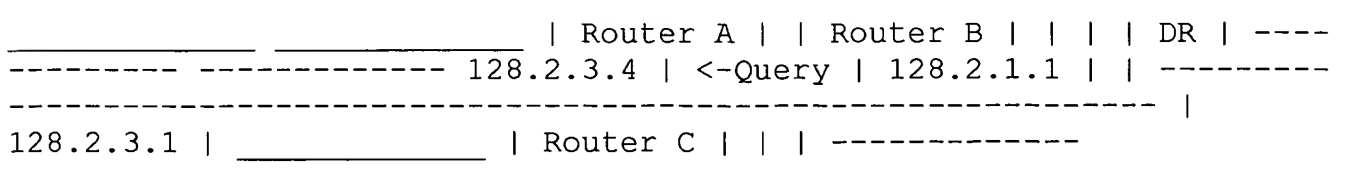


Figure 12. DVMRP Dominant Router Election

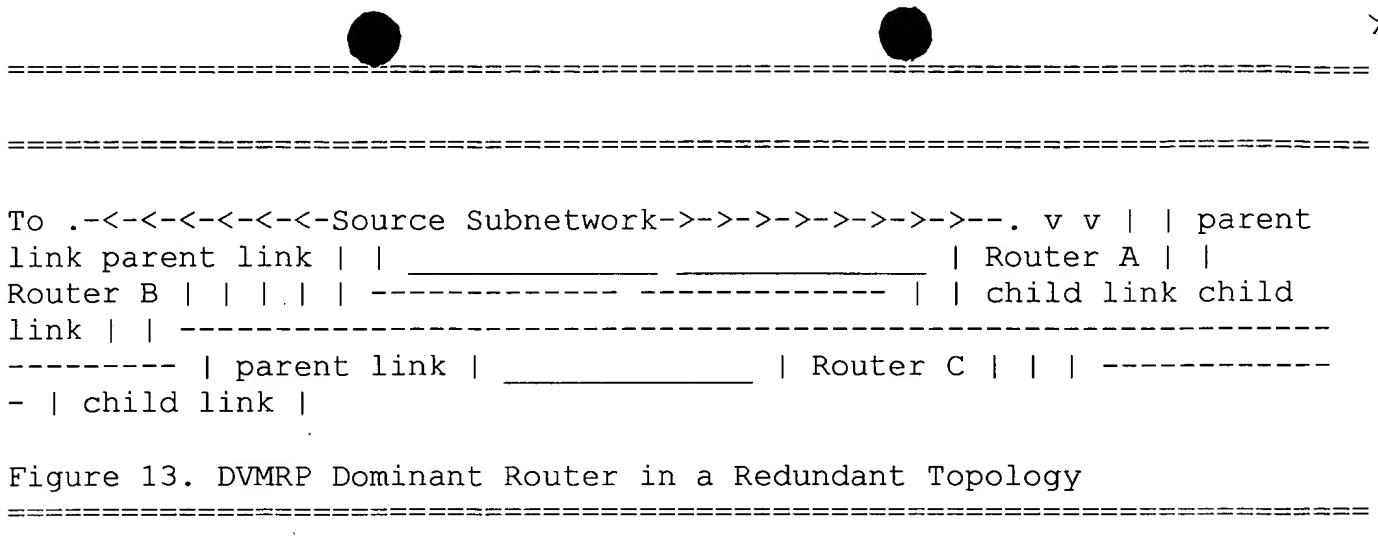


Figure 13. DVMRP Dominant Router in a Redundant Topology

7.1.4 DVMRP Routing Table

The DVMRP process periodically exchanges routing table updates with its DVMRP neighbors. These updates are logically independent of those generated by any unicast Interior Gateway Protocol.

Since the DVMRP was developed to route multicast and not unicast traffic, a router will probably run multiple routing processes in practice: One to support the forwarding of unicast traffic and another to support the forwarding of multicast traffic. (This can be convenient: A router can be configured to only route multicast IP, with no unicast IP routing. This may be a useful capability in firewalled environments.)

Consider Figure 13: There are two types of routers in this figure: dominant and subordinate; assume in this example that Router B is dominant, Router A is subordinate, and Router C is part of the downstream distribution tree. In general, which routers are dominant or subordinate may be different for each source! A subordinate router is one that is NOT on the shortest path tree back toward a source. The dominant router can tell this because the subordinate router will 'poison-reverse' the route for this source in its routing updates which are sent on the common LAN (i.e., Router A sets the metric for this source to 'infinity'). The dominant router keeps track of subordinate routers on a per-source basis...it never needs or expects to receive a prune message from a subordinate router. Only routers that are truly on the downstream distribution tree will ever need to send prunes to the dominant router. If a dominant router on a LAN has received either a poison-reversed route for a source, or prunes for all groups emanating from that source subnetwork, then it may itself send a prune upstream toward the source (assuming also that IGMP has

told it that there are no local receivers for any group from this source).

A sample routing table for a DVMRP router is shown in Figure 14. Unlike

==========

Source Subnet From Metric Status TTL Prefix Mask Gateway

128.1.0.0 255.255.0.0 128.7.5.2 3 Up 200 128.2.0.0 255.255.0.0 128.7.5.2 5 Up 150 128.3.0.0 255.255.0.0 128.6.3.1 2 Up 150 128.3.0.0 255.255.0.0 128.6.3.1 4 Up 200

Figure 14: DVMRP Routing Table

==========

the table that would be created by a unicast routing protocol such as the RIP, OSPF, or the BGP, the DVMRP routing table contains Source Prefixes and From-Gateways instead of Destination Prefixes and Next-Hop Gateways.

The routing table represents the shortest path (source-based) spanning tree to every possible source prefix in the internetwork--the Reverse Path Broadcasting (RPB) tree. The DVMRP routing table does not represent group membership or received prune messages.

The key elements in DVMRP routing table include the following items:

Source Prefix A subnetwork which is a potential or actual source of multicast datagrams.

Subnet Mask The subnet mask associated with the Source Prefix. Note that the DVMRP provides the subnet mask for each source subnetwork (in other words, the DVMRP is classless).

From-Gateway The previous-hop router leading back toward a particular Source Prefix.

TTL The time-to-live is used for table management and indicates the number of seconds before an entry is removed from the routing table.

This TTL has nothing at all to do with the TTL used in TTL-based scoping.

7.1.5 DVMRP Forwarding Table

Since the DVMRP routing table is not aware of group membership, the DVMRP process builds a forwarding table based on a combination of the information contained in the multicast routing table, known groups, and received prune messages. The forwarding table represents the local router's understanding of the shortest path source-based delivery tree for each (source, group) pair--the Reverse Path Multicasting (RPM) tree.

==========

Source Multicast TTL InPort OutPorts Prefix Group

128.1.0.0 224.1.1.1 200 1 Pr 2p3p 224.2.2.2 100 1 2p3 224.3.3.3 250 1 2 128.2.0.0 224.1.1.1 150 2 2p3

Figure 15: DVMRP Forwarding Table

==========

The forwarding table for a sample DVMRP router is shown in Figure 15. The elements in this display include the following items:

Source Prefix The subnetwork sending multicast datagrams to the specified groups (one group per row).

Multicast Group The Class D IP address to which multicast datagrams are addressed. Note that a given Source Prefix may contain sources for several Multicast Groups.

InPort The parent port for the (source, group) pair. A 'Pr' in this column indicates that a prune message has been sent to the upstream router (the From-Gateway for this Source Prefix in the DVMRP routing table).

OutPorts The child ports over which multicast datagrams for this (source, group) pair are forwarded. A 'p' in this column indicates that the router has received a prune message(s) from a (all) downstream router(s) on this port.

7.1.6 Hierarchical DVMRP (DVMRP v4.0)

The rapid growth of the MBone is placing ever-increasing demands on its routers. Essentially, today's MBone is deployed as a single, "flat" routing domain where each router is required to maintain detailed routing information to every possible subnetwork on the MBone. As the number of subnetworks continues to increase, the size of the routing tables and of the periodic update messages will continue to grow. If nothing is done about these issues, the processing and memory capabilities of the MBone routers will eventually be depleted and routing on the MBone will be degraded, or fail.

To overcome these potential scaling issues, a hierarchical version of the DVMRP is under development. In hierarchical routing, the MBone would be divided into a number of individual routing domains. Each routing domain executes its own instance of an "intra-domain" multicast routing protocol. Another protocol, or another instance of the same protocol, would be used for routing between the individual domains.

7.1.6.1 Benefits of Hierarchical Multicast Routing

Hierarchical routing reduces the demand for router resources because each router only needs to know the explicit details about routing packets to destinations within its own domain, but needs to know little or nothing about the detailed topological structure of any of the other domains. The protocol running between the domains is envisioned to maintain information about the interconnection of the domains, but not about the internal topology of each domain.

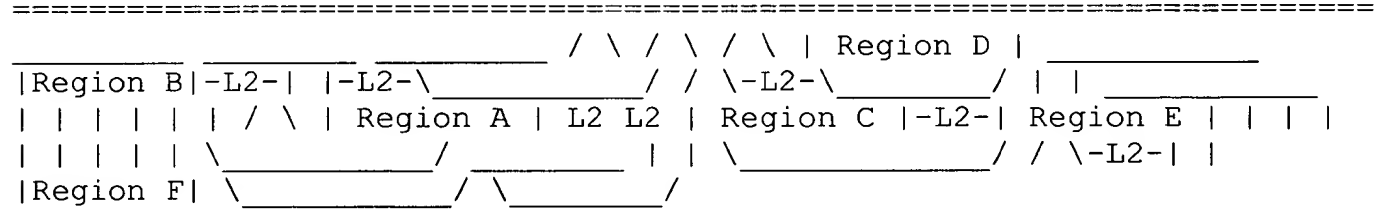


Figure 16. Hierarchical DVMRP

In addition to reducing the amount of routing information, there are several other benefits to be gained from the development and deployment of a hierarchical version of the DVMRP:

o Different multicast routing protocols may be deployed in each region of the MBone. This permits the testing and deployment of new

protocols on a domain-by-domain basis.

o The effects of an individual link or router failures are limited to only those routers operating within a single domain. Likewise, the effects of any change to the topological interconnection of regions is limited to only inter-domain routers. These enhancements are especially important when deploying a distance-vector routing protocol which can result in relatively long convergence times.

o The count-to-infinity problem associated with distance- vector routing protocols places limitations on the maximum diameter of the MBone topology. Hierarchical routing limits these diameter constraints to a single domain, instead of to the entire MBone.

7.1.6.2 Hierarchical Architecture

Hierarchical DVMRP proposes the creation of non-intersecting regions where each region has a unique Region-ID. The routers internal to a region execute any multicast routing protocols such as DVMRP, MOSPF, PIM, or CBT as a "Level 1" (L1) protocol. Each region is required to have at least one "boundary router" which is responsible for providing

inter-regional connectivity. The boundary routers execute DVMRP as a "Level 2" (L2) protocol to forward traffic between regions.

The L2 routers exchange routing information in the form of Region-IDs instead of the individual subnetwork prefixes contained within each region. With DVMRP as the L2 protocol, the inter-regional multicast delivery tree is constructed based on the (region_ID, group) pair rather than the usual (source, group) pair.

When a multicast packet originates within a region, it is forwarded according to the L1 protocol to all subnetworks containing group members. In addition, the datagram is forwarded to each of the boundary routers (L2) configured for the source region. The L2 routers tag the packet with the Region-Id and place it inside an encapsulation header for delivery to other regions. When the packet arrives at a remote region, the encapsulation header is removed before delivery to group members by the L1 routers.

7.2. Multicast Extensions to OSPF (MOSPF)

Version 2 of the Open Shortest Path First (OSPF) routing protocol is defined in RFC-1583. It is an Interior Gateway Protocol (IGP) specifically designed to distribute unicast topology information among routers belonging to a single Autonomous System. OSPF is based on link-state algorithms which permit rapid route calculation with a minimum of routing protocol traffic. In addition to efficient route calculation, OSPF is an open standard that supports hierarchical routing, load balancing, and the import of external routing information.

The Multicast Extensions to OSPF (MOSPF) are defined in RFC-1584. MOSPF routers maintain a current image of the network topology through the unicast OSPF link-state routing protocol. MOSPF enhances the OSPF protocol by providing the ability to route multicast IP traffic. The multicast extensions to OSPF are built on top of OSPF Version 2 so that a multicast routing capability can be incrementally introduced into an OSPF Version 2 routing domain. The enhancements that have been added are backwards-compatible so that routers running MOSPF will interoperate with non-multicast OSPF routers when forwarding unicast IP data traffic. Note that MOSPF, unlike DVMRP, does not provide support for tunnels.

7.2.1 Intra-Area Routing with MOSPF

Intra-Area Routing describes the basic routing algorithm employed by MOSPF. This elementary algorithm runs inside a single OSPF area and supports multicast forwarding when a source and all destination group members reside in the same OSPF area, or when the entire OSPF Autonomous System is a single area. The following discussion assumes that the reader is familiar with the basic operation of the OSPF routing protocol.

7.2.1.1 Local Group Database

Similar to the DVMRP, MOSPF routers use the Internet Group Management Protocol (IGMP) to monitor multicast group membership on directly-attached subnetworks. MOSPF routers are required to implement a "local group database" which maintains a list of directly attached groups and determines the local router's responsibility for

delivering multicast datagrams to these groups.

On any given subnetwork, the transmission of IGMP Host Membership Queries is performed solely by the Designated Router (DR). Also, the responsibility of listening to IGMP Host Membership Reports is performed only by the Designated Router (DR) and the Backup Designated Router (BDR). This means that in a mixed environment containing both MOSPF and OSPF routers, an MOSPF router must be elected the DR for the subnetwork if IGMP Queries are to be generated. This can be achieved by simply assigning all non-MOSPF routers a RouterPriority of 0 to prevent them from becoming the DR or BDR, thus allowing an MOSPF router to become the DR for the subnetwork.

The DR is responsible for communicating group membership information to all other routers in the OSPF area by flooding Group-Membership LSAs. The DR originates a separate Group-Membership LSA for each multicast group having one or more entries in the DR's local group database. Similar to Router-LSAs and Network-LSAs, Group-Membership LSAs are flooded throughout a single area only. This ensures that all remotely- originated multicast datagrams are forwarded to the specified subnetwork for distribution to local group members.

7.2.1.2 Datagram's Shortest Path Tree

The datagram's shortest path tree describes the path taken by a multicast datagram as it travels through the internetwork from the source subnetwork to each of the individual group members. The shortest path tree for each (source, group) pair is built "on demand" when a router receives the first multicast datagram for a particular (source, group) pair.

When the initial datagram arrives, the source subnetwork is located in the MOSPF link state database. The MOSPF link state database is simply the standard OSPF link state database with the addition of Group- Membership LSAs. Based on the Router- and Network-LSAs in the MOSPF link state database, a source-based shortest-path tree is constructed using Dijkstra's algorithm. After the tree is built, Group- Membership LSAs are used to prune those branches that do not lead to subnetworks containing members of this group. The output of these algorithms is a pruned source-based tree rooted at the datagram's source.

To forward multicast datagrams to downstream members of a group, each router must determine its position in the datagram's shortest path

tree. Assume that Figure 17 illustrates the shortest path tree for a particular (source, group) pair. Router E's upstream node is

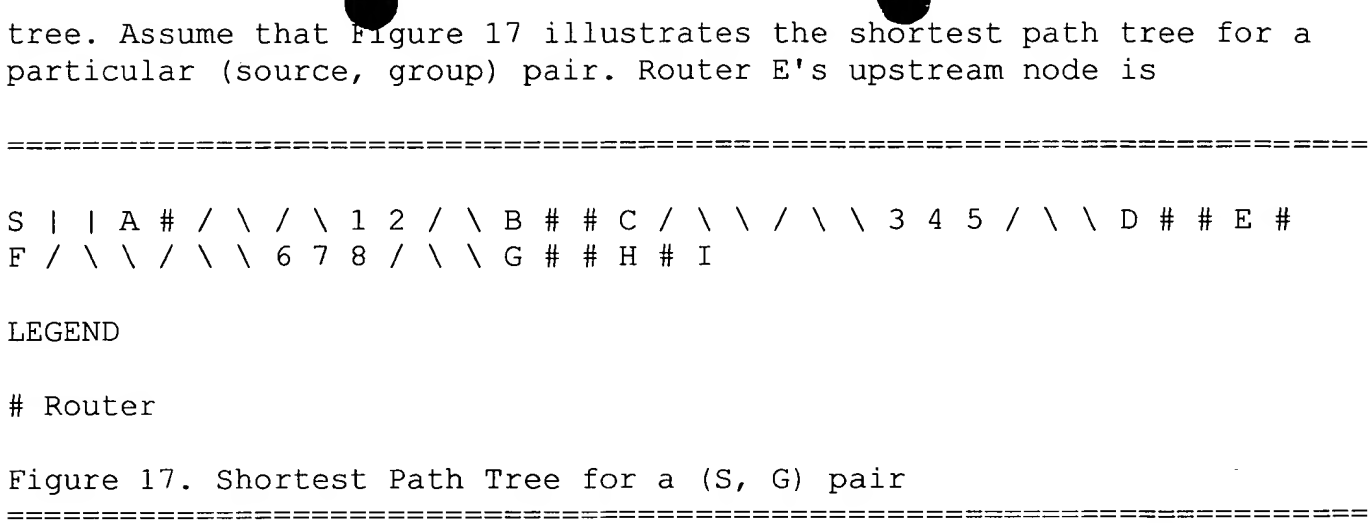


Figure 17. Shortest Path Tree for a (S, G) pair

Router B and there are two downstream interfaces: one connecting to Subnetwork 6 and another connecting to Subnetwork 7.

Note the following properties of the basic MOSPF routing algorithm:

o For a given multicast datagram, all routers within an OSPF area calculate the same source-based shortest path delivery tree. Tie-breakers have been defined to guarantee that if several equal-cost paths exist, all routers agree on a single path through the area. Unlike unicast OSPF, MOSPF does not support the concept of equal-cost multipath routing.

o Synchronized link state databases containing Group-Membership LSAs allow an MOSPF router to effectively perform the Reverse

Path Multicasting (RPM) computation "in memory." Unlike the DVMRP, this means that the first datagram of a new transmis- sion does not have to be flooded to all routers in an area.

o The "on demand" construction of the source-based delivery tree has the benefit of spreading calculations over time, resulting in a lesser impact for participating routers. Of course, this may strain the CPU(s) in a router if many (source, group) pairs appear at about the same time, or if there are a lot of events which force the router to flush and rebuild its forwarding cache. In a stable topology with long-lived multicast sessions, these effects should be minimal.

7.2.1.3 Forwarding Cache

Each MOSPF router makes its forwarding decision based on the contents

of its forwarding cache. The forwarding cache is built from the source- based shortest-path tree for each (source, group) pair and the router's local group database. After the router discovers its position in the shortest path tree, a forwarding cache entry is created containing the (source, group) pair, the upstream interface, and the downstream interface(s). At this point, all resources associated with the creation of the tree are deleted. From this point on, the forwarding cache entry is used to quickly forward all subsequent datagrams from this source to this group.

Figure 18 displays the forwarding cache for an example MOSPF router. The elements in the display include the following items:

Dest. Group The destination group address to which matching datagrams are forwarded.

Source The datagram's source host address. Each (Dest. Group, Source) pair uniquely identifies a separate forwarding cache entry.

==========

Dest. Group Source Upstream Downstream TTL

224.1.1.1 128.1.0.2 11 12 13 5 224.1.1.1 128.4.1.2 11 12 13 2
224.1.1.1 128.5.2.2 11 12 13 3 224.2.2.2 128.2.0.3 12 11 7

Figure 18: MOSPF Forwarding Cache
==========

Upstream The interface from which a matching datagram must be received

Downstream The interface(s) over which a matching datagram will be forwarded to reach known Destination group members

TTL The minimum number of hops a datagram will travel to reach the multicast group members. This allows the router to discard datagrams that do not have a high enough TTL to reach a certain group member.

The information in the forwarding cache is not aged or periodically refreshed. It is maintained as long as there are system resources

available (e.g., memory) or until the next topology change. In general, the contents of the forwarding cache will change when:

o The topology of the OSPF internetwork changes, forcing all of the shortest path trees to be recalculated. (Once the cache has been flushed, entries are not rebuilt until another packet for one of the previous (Dest. Group, Source) pairs is received.)

o There is a change in the Group-Membership LSAs indicating that the distribution of individual group members has changed.

7.2.2 Mixing MOSPF and OSPF Routers

MOSPF routers can be combined with non-multicast OSPF routers. This permits the gradual deployment of MOSPF and allows experimentation with multicast routing on a limited scale. When MOSPF and non-MOSPF routers are mixed within an Autonomous System, all routers will interoperate in the forwarding of unicast datagrams.

It is important to note that an MOSPF router is required to eliminate all non-multicast OSPF routers when it builds its source-based shortest- path delivery tree. An MOSPF router can easily determine the multicast capability of any other router based on the setting of the multicast- capable bit (MC-bit) in the Options field of each router's link state advertisements. The omission of non-multicast routers can create a number of potential problems when forwarding multicast traffic:

o The Designated Router for a multi-access network must be an MOSPF router. If a non-multicast OSPF router is elected the DR, the subnetwork will not be selected to forward multicast datagrams since a non-multicast DR cannot generate Group- Membership LSAs for its subnetwork (because it is not running IGMP, so it won't hear IGMP Host Membership Reports). To use MOSPF, it is a good idea to ensure that at least two of the MOSPF routers on each LAN have higher router_priority values

than any non-MOSPF routers. A possible strategy would be to configure any non-MOSPF routers with a router_priority of zero, so that they cannot become (B)DR.

o Multicast datagrams may be forwarded along suboptimal routes since the shortest path between two points may require traversal of a non-multicast OSPF router.

o Even though there is unicast connectivity to a destination, there may not be multicast connectivity. For example, the network may partition with respect to multicast connectivity since the only path between two points could require traversal of a non-multicast-capable OSPF router.

o The forwarding of multicast and unicast datagrams between two points may follow entirely different paths through the internetwork. This may make some routing problems a bit more challenging to debug.

7.2.3 Inter-Area Routing with MOSPF

Inter-area routing involves the case where a datagram's source and some of its destination group members reside in different OSPF areas. It should be noted that the forwarding of multicast datagrams continues to be determined by the contents of the forwarding cache which is still built from the local group database and the datagram source-based trees. The major differences are related to the way that group membership information is propagated and the way that the inter-area source-based tree is constructed.

7.2.3.1 Inter-Area Multicast Forwarders

In MOSPF, a subset of an area's Area Border Routers (ABRs) function as "inter-area multicast forwarders." An inter-area multicast forwarder is responsible for the forwarding of group membership information and multicast datagrams between areas. Configuration parameters determine whether or not a particular ABR also functions as an inter-area multicast forwarder.

Inter-area multicast forwarders summarize their attached areas' group membership information to the backbone by originating new Group-Membership LSAs into the backbone area. It is important to note that the summarization of group membership in MOSPF is asymmetric. This means that group membership information from non-backbone areas is flooded into the backbone. However, group membership from the backbone or from other non-backbone areas is not flooded into any non-backbone area(s).

To permit the forwarding of multicast traffic between areas, MOSPF introduces the concept of a "wild-card multicast receiver." A wild-card multicast receiver is a router that receives all multicast traffic

generated in an area, regardless of the multicast group membership.

In non-backbone areas, all inter-area multicast forwarders operate as wild-card multicast receivers. This guarantees that all multicast traffic originating in a non-backbone area is delivered to its inter-area multicast forwarder, and then if necessary into the backbone area.

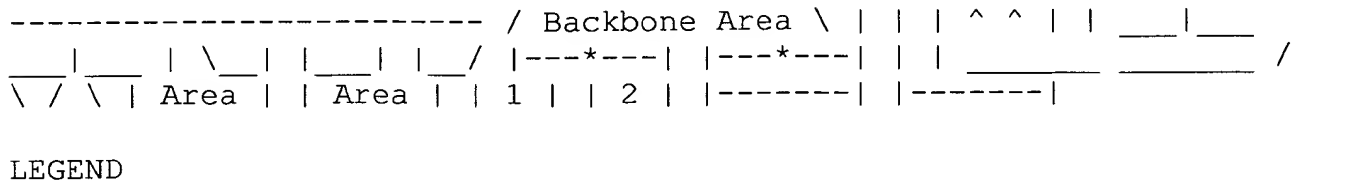


Figure 19. Inter-Area Routing Architecture

Since the backbone has group membership knowledge for all areas, the datagram can then be forwarded to group members residing in the backbone and other non-backbone areas. The backbone area does not require wild-card multicast receivers because the routers in the backbone area have complete knowledge of group membership information for the entire OSPF system.

7.2.3.2 Inter-Area Datagram Shortest-Path Tree

In the case of inter-area multicast routing, it is often impossible to build a complete datagram shortest-path delivery tree. Incomplete trees are created because detailed topological and group membership information for each OSPF area is not distributed between OSPF areas. To overcome these limitations, topological estimates are made through the use of wild-card receivers and OSPF Summary-Links LSAs.

There are two cases that need to be considered when constructing an inter-area shortest-path delivery tree. The first involves the condition when the source subnetwork is located in the same area as the router performing the calculation. The second situation occurs when the

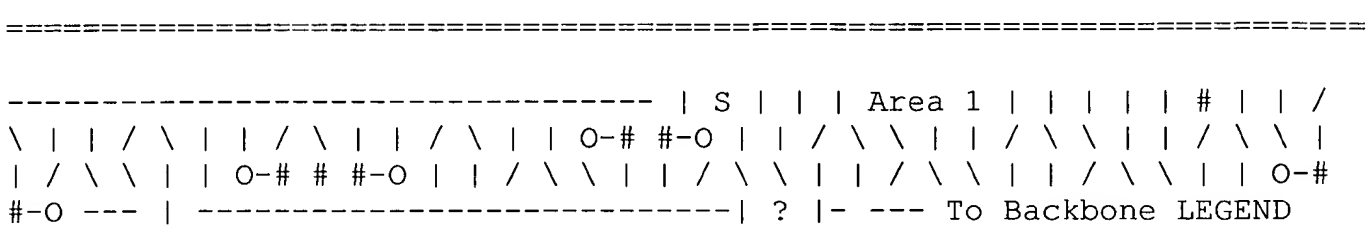


Figure 20. Datagram Shortest Path Tree (Source in Same Area)

source subnetwork is located in a different area than the router performing the calculation.

If the source of a multicast datagram resides in the same area as the router performing the calculation, the pruning process must be careful to ensure that branches leading to other areas are not removed from the tree. Only those branches having no group members nor wild-card multicast receivers are pruned. Branches containing wild-card multicast receivers must be retained since the local routers do not know if there are group members residing in other areas.

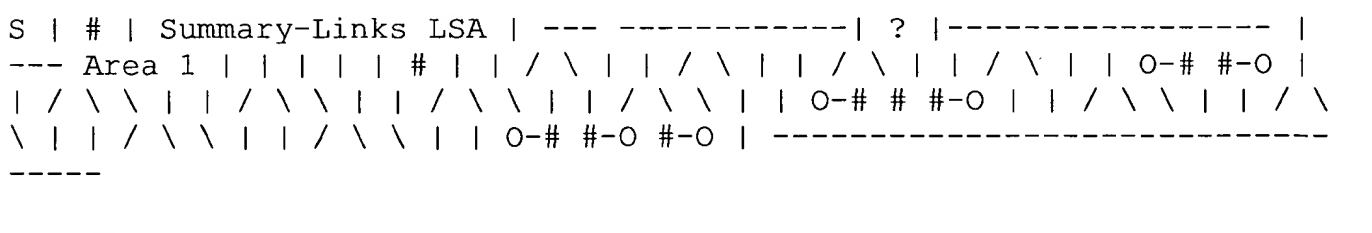


Figure 21. Shortest Path Tree (Source in Different Area)

If the source of a multicast datagram resides in a different area than the router performing the calculation, the details describing the local topology surrounding the source station are not known. However, this information can be estimated using information provided by Summary-Links LSAs for the source subnetwork. In this case, the base of the tree begins with branches directly connecting the source subnetwork to each of the local area's inter-area multicast forwarders. The inter-area multicast forwarders must be included in the tree since any multicast datagrams originating outside the local area will enter the area via an inter-area multicast forwarder.

Since each inter-area multicast forwarder is also an ABR, it must maintain a separate link state database for each attached area. This means that each inter-area multicast forwarder is required to calculate a separate forwarding tree for each of its attached areas. After the individual trees are calculated, they are merged into a single forwarding cache entry for the (source, group) pair and then the individual trees are discarded.

7.2.4 Inter-Autonomous System Multicasting with MOSPF

Inter-Autonomous System multicasting involves the situation where a datagram's source and at least some of its destination group members reside in different OSPF Autonomous Systems. It should be emphasized that in OSPF terminology "inter-AS" communication also refers to connectivity between an OSPF domain and another routing domain which could be within the same Autonomous System from the perspective of an Exterior Gateway Protocol.

To facilitate inter-AS multicast routing, selected Autonomous System Boundary Routers (ASBRs) are configured as "inter-AS multicast forwarders." MOSPF makes the assumption that each inter-AS multicast forwarder executes an inter-AS multicast routing protocol (e.g., DVMRP) which forwards multicast datagrams in a reverse path forwarding (RPF) manner. Each inter-AS multicast forwarder functions as a wild-card multicast receiver in each of its attached areas. This guarantees that each inter-AS multicast forwarder remains on all pruned shortest-path trees and receives all multicast datagrams, regardless of the multicast group membership.

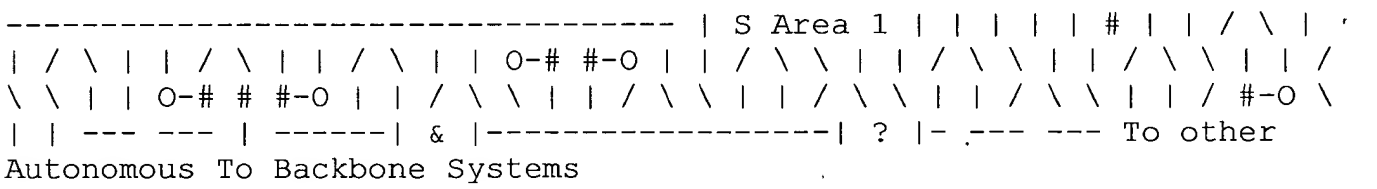
Three cases need to be considered when describing the construction of an inter-AS shortest-path delivery tree. The first occurs when the source subnetwork is located in the same area as the router

performing the calculation. For the second case, the source subnetwork resides in a different area than the router performing the calculation. The final case occurs when the source subnetwork is located in a different AS than the router performing the calculation.

The first two cases are similar to the inter-area examples described in the previous section. The only enhancement is that inter-AS multicast forwarders must also be included on the pruned shortest path delivery

tree. Branches containing inter-AS multicast forwarders must be retained since the local routers do not know if there are group members residing in other Autonomous Systems. When a multicast datagram arrives at an inter-AS multicast forwarder, it is the responsibility of the ASBR to determine whether the datagram should be forwarded outside of the local Autonomous System.

Figure 22 illustrates a sample inter-AS shortest path delivery tree when the source subnetwork resides in the same area as the router performing the calculation.



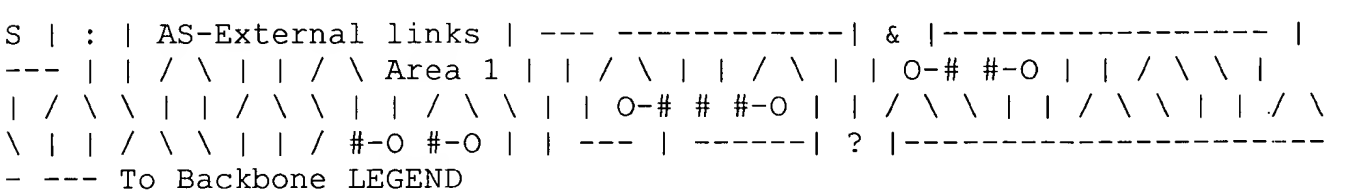
LEGEND

S Source Subnetwork O Subnet Containing Group Members # Intra-Area MOSPF Router ? Inter-Area Multicast Forwarder & Inter-AS Multicast Forwarder

Figure 22. Inter-AS Datagram Shortest Path Tree (Source in Same Area)

If the source of a multicast datagram resides in a different Autonomous System than the router performing the calculation, the details describing the local topology surrounding the source station

are not known. However, this information can be estimated using the multicast- capable AS-External Links describing the source subnetwork. In this case, the base of the tree begins with branches directly connecting the source subnetwork to each of the local area's inter-AS multicast forwarders.



S Source Subnetwork O Subnet Containing Group Members # Intra-Area MOSPF Router ? Inter-Area Multicast Forwarder & Inter-AS Multicast Forwarder

Figure 23. Inter-AS Datagram Shortest Path Tree (Source in Different AS)

Figure 23 shows a sample inter-AS shortest-path delivery tree when the inter-AS multicast forwarder resides in the same area as the router performing the calculation. If the inter-AS multicast forwarder is located in a different area than the router performing the calculation, the topology surrounding the source is approximated by combining the Summary-ASBR Link with the multicast capable AS-External Link.

As a final point, it is important to note that AS External Links are not imported into Stub areas. If the source is located outside of the stub area, the topology surrounding the source is estimated by the Default Summary Links originated by the stub area's intra-area multicast forwarder rather than the AS-External Links.

7.3 Protocol-Independent Multicast (PIM)

The Protocol Independent Multicast (PIM) routing protocol is currently under development by the Inter-Domain Multicast Routing (IDMR) working group of the IETF. The objective of the IDMR working group is to develop one--or possibly more than one--standards-track multicast routing protocol(s) that can provide scaleable inter-domain multicast routing across the Internet.

PIM receives its name because it is not dependent on the mechanisms provided by any particular unicast routing protocol. However, any implementation supporting PIM requires the presence of a unicast routing protocol to provide routing table information and to adapt to topology changes.

PIM makes a clear distinction between a multicast routing protocol that is designed for dense environments and one that is designed for sparse environments. Dense-mode refers to a protocol that is designed to operate in an environment where group members are relatively densely packed and bandwidth is plentiful. Sparse-mode refers to a protocol that is optimized for environments where group members are distributed across many regions of the Internet and bandwidth is not necessarily widely available. It is important to note that sparse-mode does not imply that the group has a few members, just that they are widely dispersed across the Internet.

The designers of PIM argue that DVMRP and MOSPF were developed for environments where group members are densely distributed. They emphasize that when group members and senders are sparsely distributed across a wide area, DVMRP and MOSPF do not provide the most efficient multicast delivery service. DVMRP periodically sends multicast packets over many links that do not lead to group members, while MOSPF can send group membership information over links that do not lead to senders or receivers.

7.3.1 PIM-Dense Mode (PIM-DM)

While the PIM architecture was driven by the need to provide scaleable sparse-mode delivery trees, PIM also defines a new dense-mode protocol instead of relying on existing dense-mode protocols such as DVMRP and MOSPF. It is envisioned that PIM-DM would be deployed in resource rich environments, such as a campus LAN where group membership is relatively dense and bandwidth is likely to be readily available. PIM-DM's control messages are similar to PIM-SM's by design.

PIM - Dense Mode (PIM-DM) is similar to DVMRP in that it employs the Reverse Path Multicasting (RPM) algorithm. However, there are several important differences between PIM-DM and DVMRP:

o To find routes back to sources, PIM-DM relies on the presence of an existing unicast routing table. PIM-DM is independent of the mechanisms of any specific unicast routing protocol. In contrast, DVMRP contains an integrated routing protocol that makes use of its own RIP-like exchanges to build its own unicast routing table (so a router may orient itself with respect to active source(s). MOSPF augments the information in the OSPF link state database, thus MOSPF must run in conjunction with OSPF.

o Unlike the DVMRP which calculates a set of child interfaces for each (source, group) pair, PIM-DM simply forwards multicast traffic on all downstream interfaces until explicit prune messages are received. PIM-DM is willing to accept packet duplication to eliminate routing protocol dependencies and to avoid the overhead inherent in determining the parent/child relationships.

For those cases where group members suddenly appear on a pruned branch of the delivery tree, PIM-DM, like DVMRP, employs graft messages to re-attach the previously pruned branch to the delivery tree.

8. SHARED TREE ("SPARSE MODE") ROUTING PROTOCOLS

The most recent additions to the set of multicast routing proto- cols are based on a shared delivery tree.

These emerging routing protocols include: o Protocol Independent Multicast - Sparse Mode (PIM-SM), and o Core-Based Trees (CBT). Each of these routing protocols is designed to operate efficiently over a wide area network where bandwidth is scarce and group members may

be sparsely distributed. Their ultimate goal is to provide scaleable interdomain multicast routing across the Internet.

8.1 Protocol-Independent Multicast - Sparse Mode (PIM-SM)

As described previously, PIM also defines a "dense-mode" or source-

based tree variant. The two protocols are quite unique, and other than control messages, they have very little else in common. Because PIM integrates control message processing and data packet forwarding among PIM-Sparse and -Dense Modes, a single PIM router can run different modes for different groups, as desired.

PIM-Sparse Mode (PIM-SM) is being developed to provide a multicast routing protocol that provides efficient communication between members of sparsely distributed groups--the type of groups that are likely to be common in wide-area internetworks. PIM's designers observe that several hosts wishing to participate in a multicast conference do not justify flooding the entire internetwork periodically with the group's multicast traffic.

Noting today's existing MBone scaling problems, and extrapolating to a future of ubiquitous multicast (overlaid with perhaps thousands of small, widely dispersed groups), it is not hard to imagine that existing multicast routing protocols will experience scaling problems. To eliminate these potential scaling issues, PIM-SM is designed to limit multicast traffic so that only those routers interested in receiving traffic for a particular group "see" it.

PIM-SM differs from existing dense-mode protocols in two key ways:

o Routers with adjacent or downstream members are required to explicitly join a sparse mode delivery tree by transmitting join messages. If a router does not join the pre-defined delivery tree, it will not receive multicast traffic addressed to the group.

In contrast, dense-mode protocols assume downstream group membership and forward multicast traffic on downstream links until explicit prune messages are received. Thus, the default forwarding action of dense-mode routing protocols is to forward all traffic, while the default action of a sparse-mode protocol is to block traffic unless it has been explicitly requested.

o PIM-SM evolved from the Core-Based Trees (CBT) approach in that it employs the concept of a "core" (or rendezvous point (RP) in PIM-SM terminology) where receivers "meet" sources. The creator of each multicast group selects a primary RP and a small set of alternative RPs, known as the RP-set. For each group, there is only a single active RP (which is uniquely determined by a hash function).

==========

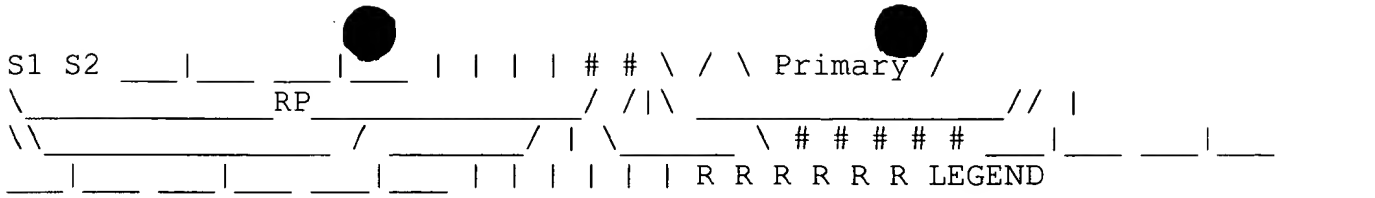


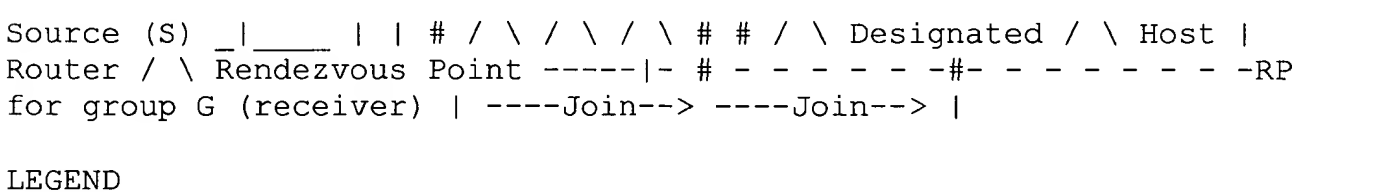
Figure 24 Primary Rendezvous Point

When joining a group, each receiver uses IGMP to notify its directly-attached router, which in turn joins the multicast delivery tree by sending an explicit PIM-Join message hop-by-hop toward the group's primary RP. A source uses the RP to announce its presence, and act as a conduit to members that have joined the group. This model requires sparse-mode routers to maintain a bit of state (i.e., the RP-set for each defined sparse-mode group) prior to the arrival of data. In contrast, dense mode protocols are data-driven, since they do not store any state for a group until the arrival of the first data packet.

8.1.1 Directly Attached Host Joins a Group

When there is more than one PIM router connected to a multi-access LAN, the router with the highest IP address is selected to function as the Designated Router (DR) for the LAN. The DR may or may not be responsible for the transmission of IGMP Host Membership Query messages, but does send Join/Prune messages toward the RP, and maintains the status of the active RP for local senders to multicast groups.

When the DR receives an IGMP Report message for a new group, the DR determines if the group is RP-based or not by examining the group address. If the address indicates a SM group (by virtue of the group-specific state that even inactive groups have stored in all PIM routers), the DR performs a deterministic hash function over the



PIM Router RP Rendezvous Point

Figure 25: Host Joins a Multicast Group

group's RP-set to uniquely determine the primary RP for the group. (Otherwise, this is a dense-mode group and dense-mode forwarding rules apply.)

After performing the lookup, the DR creates a multicast forwarding cache entry for the (*, group) pair and transmits a unicast PIM-Join message toward the primary RP for this specific group. The (*, group) notation indicates an (any source, group) pair. The intermediate routers forward the unicast PIM-Join message, creating a forwarding cache entry for the (*, group) pair only if such a forwarding entry does not yet exist. Intermediate routers must create a forwarding cache entry so that they will be able to forward future traffic downstream toward the DR which originated the PIM-Join message.

8.1.2 Directly Attached Source Sends to a Group

When a source first transmits a multicast packet to a group, its DR forwards the datagram to the primary RP for subsequent distribution along the group's delivery tree. The DR encapsulates the initial multicast packets in a PIM-SM-Register packet and unicasts them toward the primary RP for the group. The PIM-SM-Register packet informs the RP of a new source which causes the active RP to transmit PIM-Join messages back toward the source's DR. The routers between the RP and the source's DR use the re- ceived PIM-Join messages (from the RP) to

create forwarding state for the new (source, group) pair. Now all routers from the active RP for this sparse-mode group to the source's DR will be able to forward future unencapsulated multicast packets from this source subnetwork to the RP. Until the (source, group) state has been created in all the routers between the RP and source's DR, the DR must continue to send the source's multicast IP packets to the RP as unicast packets encapsulated within unicast PIM-Register packets. The DR may stop forwarding multicast packets encapsulated in this manner once it has received a PIM-Register-Stop message from the active RP for this group. The RP may send PIM-Register-Stop messages if there are no downstream receivers for a group, or if the RP has successfully joined the (source, group) tree (which originates at the source's DR).

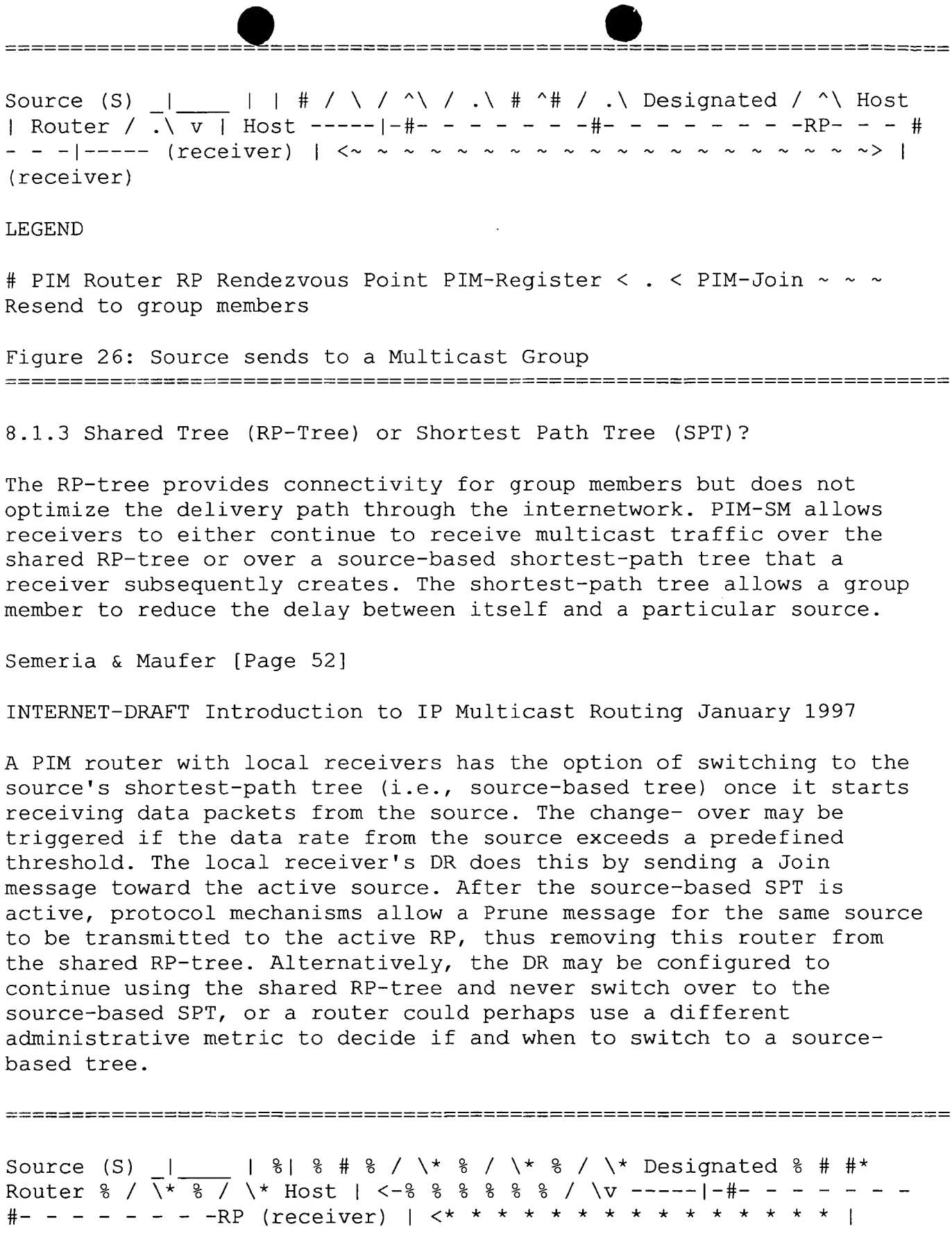


Figure 26: Source sends to a Multicast Group

8.1.3 Shared Tree (RP-Tree) or Shortest Path Tree (SPT)?

The RP-tree provides connectivity for group members but does not optimize the delivery path through the internetwork. PIM-SM allows receivers to either continue to receive multicast traffic over the shared RP-tree or over a source-based shortest-path tree that a receiver subsequently creates. The shortest-path tree allows a group member to reduce the delay between itself and a particular source.

A PIM router with local receivers has the option of switching to the source's shortest-path tree (i.e., source-based tree) once it starts receiving data packets from the source. The change- over may be triggered if the data rate from the source exceeds a predefined threshold. The local receiver's DR does this by sending a Join message toward the active source. After the source-based SPT is active, protocol mechanisms allow a Prune message for the same source to be transmitted to the active RP, thus removing this router from the shared RP-tree. Alternatively, the DR may be configured to continue using the shared RP-tree and never switch over to the source-based SPT, or a router could perhaps use a different administrative metric to decide if and when to switch to a source-based tree.

PIM Router RP Rendezvous Point * * RP Tree % % SPT Tree

Figure 27: Shared RP-Tree and Shortest Path Tree (SPT)

8.1.4 Unresolved Issues

It is important to note that PIM is an Internet draft. This means that it is still early in its development cycle and clearly a "work in

progress." There are several important issues that require further research, engineering, and/or experimentation:

o PIM-SM requires routers to maintain a non-trivial amount of state information to describe sources and groups.

o Some multicast routers will be required to have both PIM interfaces and non-PIM interfaces. The interaction and sharing of multicast routing information between PIM and other multicast routing protocols is still being defined.

Due to these reasons, especially the need to get operational experience with the protocol, when PIM is finally published as an RFC, it will not immediately be placed on the standards-track; rather it will be classified as experimental. After sufficient operational experience has been obtained, presumably a slightly altered specification will be defined that incorporates lessons learned during the experimentation phase, and that new specification will then be placed on the standards track.

8.2 Core-Based Trees (CBT)

Core Based Trees is another multicast architecture that is based on a shared delivery tree. It is specifically intended to address the important issue of scalability when supporting multicast applications across the public Internet. CBT is also designed to enable

interoperability between distinct "clouds" on the Internet, each executing a different multicast routing protocol.

Similar to PIM, CBT is protocol-independent. CBT employs the information contained in the unicast routing table to build its shared delivery tree. It does not care how the unicast routing table is derived, only that a unicast routing table is present. This feature allows CBT to be deployed without requiring the presence of any specific unicast routing protocol. 8.2.1 Joining a Group's Shared Tree When a multi-access network has more than one CBT router, one of the routers is elected the designated router (DR) for the subnetwork. The DR is responsible for transmitting IGMP Queries and for initiating the construction of a branch that links directly-attached group members to the shared distribution tree for the group. The router on the subnetwork with the lowest IP address is elected the IGMP Querier and also serves as the CBT DR.

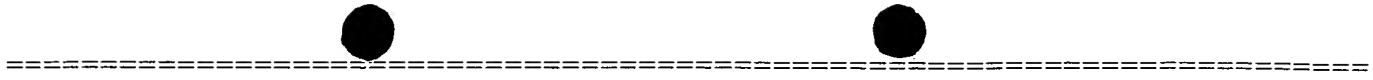
When the DR receives an IGMP Host Membership Report for a new group, it transmits a CBT Join-Request to the next-hop router on the unicast path

to the "target core" for the multicast group. The identification of the "target core" is based on static configuration.

The Join-Request is processed by all intermediate CBT routers, each of which identifies the interface on which the Join-Request was received as part of this group's delivery tree. The intermediate routers continue to forward the Join-Request toward the target core and to mark local interfaces until the request reaches either 1) a core router, or 2) a router that is already on the distribution tree for this group.

In either case, this router stops forwarding the Join-Request and responds with a Join-Ack which follows the path back to the DR which initiated the Join-Request. The Join-Ack fixes the state in each of the intermediate routers causing the interfaces to become part of the distribution tree for the multicast group. The newly constructed branch is made up of non-core (i.e., "on-tree") routers providing the shortest path between a member's directly attached DR and a core.

Once a branch is created, each child router monitors the status of its parent router with a keepalive mechanism. A child router periodically unicasts a CBT-Echo-Request to its parent router which is then required to respond with a unicast CBT-Echo-Reply message.

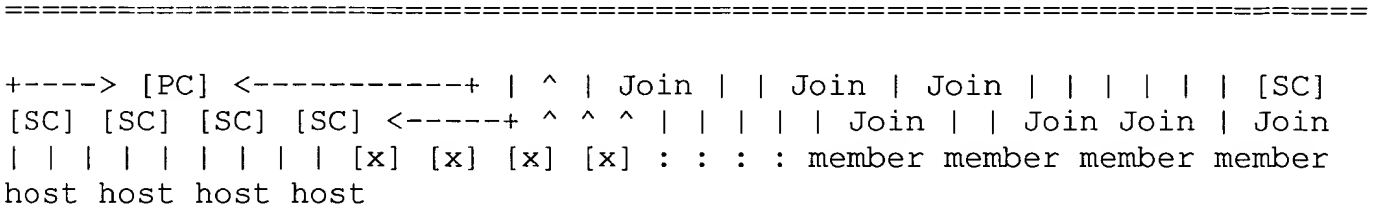


It is only necessary to implement a single "keepalive" mechanism on each link regardless of the number of multicast groups that are sharing the link. If for any reason the link between the child and parent should fail, the child is responsible for re-attaching itself and its downstream children to the shared delivery tree.

8.2.2 Primary and Secondary Cores

Instead of a single active "core" or "rendezvous point," CBT may have multiple active cores to increase robustness. The initiator of a multicast group elects one of these routers as the Primary Core, while all other cores are classified as Secondary Cores. The Primary Core must be uniquely identified for the entire multi- cast group.

Whenever a group member joins to a secondary core, the secondary core router ACKs the Join-Request and then joins toward the Primary Core. Since each Join-Request contains the identity of the Primary Core for the group, the secondary core can easily determine the identity of the Primary Core for the group. This simple process allows the CBT tree to become fully connected as individual members join the multicast group.



LEGEND

[PC] Primary Core Router [SC] Secondary Core Router [x] Member-hosts'

directly-attached routers

Figure 29: Primary and Secondary Core Routers

8.2.3 Data Packet Forwarding

After a Join-Ack is received by an intermediate router, it creates a CBT forwarding information base (FIB) entry listing all interfaces that are part of the specified group's delivery tree. When a CBT router receives a packet addressed to the multicast group, it simply forwards the packet over all outgoing interfaces as specified by the FIB entry for the group.

A CBT router may forward a multicast data packet in either "CBT Mode" or "Native Mode."

o CBT Mode is designed for operation in heterogeneous environments that may include non-multicast capable routers or mrouters that do not implement (or are not configured for) CBT. Under these conditions, CBT Mode is used to encapsulate the data packet in a CBT header and "tunnel" it between CBT-capable routers (or islands).

o Native Mode is designed for operation in a homogeneous environment where all routers implement the CBT routing protocol and no specialized encapsulation is required.

8.2.4 Non-Member Sending

Similar to other multicast routing protocols, CBT does not require that the source of a multicast packet be a member of the multicast group. However, for a multicast data packet to reach the core tree for the group, at least one CBT-capable router must be present on the non-member source station's subnetwork. The local CBT-capable router employs CBT Mode encapsulation and unicasts the data packet toward a core for the multicast group. When the encapsulated packet encounters an on-tree router (or the target core), the packet is forwarded as required by the CBT specification.

8.2.5 Emulating Shortest-Path Trees

The most common criticism of shared tree protocols is that they offer sub-optimal routes and that they create high levels of traffic concentration at the core routers. One recent proposal in CBT technology is a mechanism to dynamically reconfigure the core-based

tree so that it becomes rooted at the source station's local CBT router. In effect, the CBT becomes a source-based tree but still remains a CBT (one with a core that now happens to be adjacent to the source). If successfully tested and demonstrated, this technique could allow CBT to emulate a shortest-path tree, providing more-optimal routes and reducing traffic concentration among the cores. These new mechanisms are being designed with an eye toward preserving CBT's simplicity and scalability, while addressing key perceived weaknesses of the CBT protocol. Note that PIM-SM also has a similar technique whereby a source-based delivery tree can be selected by certain receivers.

For this mechanism, every CBT router is responsible for monitoring the transmission rate and duration of each source station on a directly attached subnetwork. If a pre-defined threshold is exceeded, the local CBT router may initiate steps to transition the CBT tree so that the group's receivers become joined to a "core" that is local to the source station's subnetwork. This is accomplished by having the local router encapsulate traffic in CBT Mode and place its own IP address in the "first-hop router" field. All routers on the CBT tree examine the "first-hop router" field in every CBT Mode data packet. If this field contains a non-NULL value, each router transmits a Join-Request toward the address specified in the "first-hop router" field. It is important to note that on the publication date of this "Introduction to IP Multicast Routing" RFC, these proposed mechanisms to support dynamic source-migration of cores have not yet been tested, simulated, or demonstrated.

8.2.6 CBT Multicast Interoperability

Multicast interoperability is being defined in several stages. Stage 1 is concerned with the attachment of non-DVMRP stub domains to a DVMRP backbone (e.g., the MBone). Work is currently underway in the IDMR working group to describe the attachment of stub-CBT and stub-PIM domains to a DVMRP backbone. The next stage will focus on developing methods of connecting non-DVMRP transit domains to a DVMRP backbone.

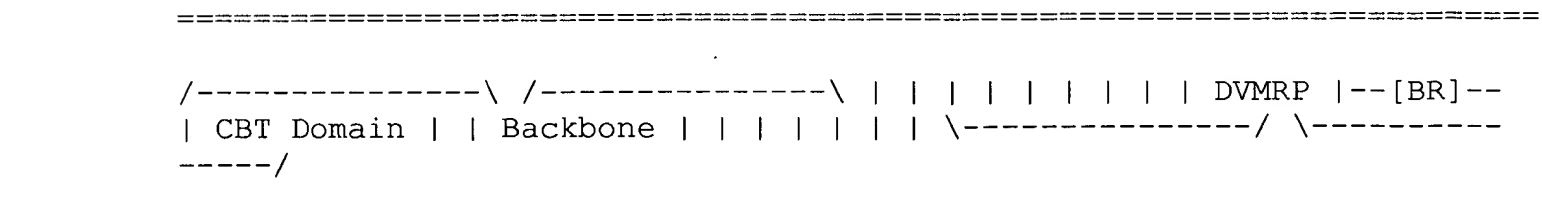


Figure 30: Domain Border Routers (BRs)

CBT interoperability will be achieved through the deployment of domain border routers (BRs) which enable the forwarding of multicast traffic between the CBT and DVMRP domains. The BR implements DVMRP and CBT on different interfaces and is responsible for forwarding data across the domain boundary.

The BR is also responsible for exporting selected routes out of the CBT domain into the DVMRP domain. While the CBT domain never needs to import routes, the DVMRP backbone needs to import routes to sources of

traffic from within the CBT domain. The routes must be imported so that DVMRP can perform the RPF check (which is required for construction of its forwarding table).

9. REFERENCES

9.1 Requests for Comments (RFCs)

1075 "Distance Vector Multicast Routing Protocol," D. Waitzman, C. Partridge, and S. Deering, November 1988.

1112 "Host Extensions for IP Multicasting," Steve Deering, August 1989.

1583 "OSPF Version 2," John Moy, March 1994.

1584 "Multicast Extensions to OSPF," John Moy, March 1994.

1585 "MOSPF: Analysis and Experience," John Moy, March 1994.

1700 "Assigned Numbers," J. Reynolds and J. Postel, October 1994. (STD 2)

1800 "Internet Official Protocol Standards," Jon Postel, Editor, July 1995.

1812 "Requirements for IP version 4 Routers," Fred Baker, Editor, June 1995

9.2 Internet Drafts

"Core Based Trees (CBT) Multicast: Architectural Overview," <draft-

ietf-idmr-cbt-arch-03.txt>, A. J. Ballardie, September 19, 1996.

"Core Based Trees (CBT) Multicast: Protocol Specification," <draft-ietf-idmr-cbt-spec-06.txt>, A. J. Ballardie, November 21, 1995.

"Hierarchical Distance Vector Multicast Routing for the MBone," Ajit Thyagarajan and Steve Deering, July 1995.

"Internet Group Management Protocol, Version 2," <draft-ietf- idmr-igmp-v2-05.txt>, William Fenner, October 25, 1996.

"Internet Group Management Protocol, Version 3," <draft-cain- igmp-00.txt>, Brad Cain, Ajit Thyagarajan, and Steve Deering, Expires March 8, 1996.

"Protocol Independent Multicast (PIM): Motivation and Architecture," <draft-ietf-idmr-pim-arch-04.ps>, S. Deering, D. Estrin, D. Farinacci, V. Jacobson, C. Liu, and L. Wei, September 11, 1996.

"Protocol Independent Multicast (PIM), Dense Mode Protocol Specification," <draft-ietf-idmr-pim-dm-spec-04.ps>, D. Estrin, D. Farinacci, V. Jacobson, C. Liu, L. Wei, P. Sharma, and A. Helmy, September 16, 1996.

"Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification," <draft-ietf-idmr-pim-sm-spec-09.ps>, S. Deering, D. Estrin, D. Farinacci, V. Jacobson, C. Liu, L. Wei, P. Sharma, and A Helmy, September 19, 1996.

9.3 Textbooks

Comer, Douglas E. Internetworking with TCP/IP Volume 1 Principles, Protocols, and Architecture Second Edition, Prentice Hall, Inc. Englewood Cliffs, New Jersey, 1991

Huitema, Christian. Routing in the Internet, Prentice Hall, Inc. Englewood Cliffs, New Jersey, 1995

Stevens, W. Richard. TCP/IP Illustrated: Volume 1 The Protocols, Addison Wesley Publishing Company, Reading MA, 1994

Wright, Gary and W. Richard Stevens. TCP/IP Illustrated: Volume 2 The Implementation, Addison Wesley Publishing Company, Reading MA, 1995

9.4 Other

Deering, Steven E. "Multicast Routing in a Datagram Internetwork," Ph.D. Thesis, Stanford University, December 1991.

Ballardie, Anthony J. "A New Approach to Multicast Communication in a Datagram Internetwork," Ph.D. Thesis, University of London, May 1995.

10. SECURITY CONSIDERATIONS

Security issues are not discussed in this memo.

11. AUTHORS' ADDRESSES

Chuck Semeria 3Com Corporation 5400 Bayfront Plaza P.O. Box 58145 Santa Clara, CA 95052-8145

Phone: +1 408 764-7201 Email: <Chuck_Semeria@3Com.com>

Tom Maufer 3Com Corporation 5400 Bayfront Plaza P.O. Box 58145 Santa Clara, CA 95052-8145

Phone: +1 408 764-8814 Email: <maufer@3Com.com>